



Ingeniare. Revista Chilena de Ingeniería

ISSN: 0718-3291

facing@uta.cl

Universidad de Tarapacá

Chile

Mejia, Marcela; Rubiano, Astrid; Alzate, Marco  
Segmentación de imágenes de células cervicales y evaluación de características para  
detección de lesiones neoplásicas  
Ingeniare. Revista Chilena de Ingeniería, vol. 24, núm. 2, abril, 2016, pp. 274-289  
Universidad de Tarapacá  
Arica, Chile

Disponible en: <http://www.redalyc.org/articulo.oa?id=77245711010>

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

redalyc.org

Sistema de Información Científica  
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal  
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

## Segmentación de imágenes de células cervicales y evaluación de características para detección de lesiones neoplásicas

### *Cervix cells image segmentation and evaluation of features for cervical neoplasia detection*

Marcela Mejía<sup>1</sup>    Astrid Rubiano<sup>1</sup>    Marco Alzate<sup>2</sup>

Recibido 10 de enero de 2014, aceptado 5 de noviembre de 2015

*Received: January 10, 2014    Accepted: November 5, 2015*

#### RESUMEN

El cáncer cervical puede ser curado si se detecta y trata oportunamente, para ello la prueba Pap ha sido fundamental. En este contexto, una ayuda tecnológica puede reducir la naturaleza subjetiva del diagnóstico, pero existen dificultades en su construcción. Aquí abordamos dos de ellas: la identificación del citoplasma y el núcleo de cada célula, y la determinación de un conjunto de características relevantes para la detección de lesiones neoplásicas. En este artículo presentamos dos aportes. Primero se propone un método interactivo de segmentación basado en procesamiento morfológico multiespectral en el que las imperfecciones más engañosas de las imágenes son eliminadas con una interacción simple del analista. Segundo, se hace un análisis de la relevancia de algunas variables que caracterizan los tamaños relativos del núcleo y el citoplasma, sus formas, sus texturas y la rugosidad de sus bordes. El análisis se basa en las medidas de desempeño de un detector que utiliza extracción de características mediante análisis de componentes principales (PCA) y separación de células normales y lesionadas mediante una máquina de vectores de soporte (SVM). Encontramos que una mínima interacción con el médico permite obtener segmentaciones mucho más precisas y confiables. De otro lado, encontramos que las características más relevantes para detección de lesiones neoplásicas son los tamaños relativos del núcleo y del citoplasma y sus formas, mientras que otras características, como la textura y la rugosidad, son menos relevantes.

Palabras clave: Cáncer cervical, prueba Pap, análisis morfológico, análisis multiespectral, PCA, SVM.

#### ABSTRACT

*Cervical cancer can be cured if detected and treated early, for which the Pap test has been fundamental. In this context, technological aids can reduce the subjective nature of the diagnosis, although there are still several issues to be solved. Here we address two of them: the identification of the cytoplasm and the nucleus in a cell image, and the determination of a set of relevant features for the detection of injured cells. In this paper we present two contributions. First, we propose an interactive segmentation method based on multispectral morphological processing in which most misleading imperfections are eliminated with a simple interaction of the analyst. Second, we made an analysis of the relevance of certain variables that characterize the relative sizes of nucleus and cytoplasm, their shapes, textures and roughness of the edges. The analysis is based on performance measures of a detector that uses feature extraction through principal component analysis (PCA) and separation of normal and injured cells by a support vector machine (SVM). We have found that minimal interaction with the physician allows for much more accurate and reliable segmentation than purely automatic methods. On the other hand, we have found that the most important characteristics for detection of injured cells are the relative sizes of the nucleus and the cytoplasm and their form, while other features, such as texture and roughness, are less relevant.*

*Keywords: Cervical cancer, Pap test, morphological analysis, multispectral analysis, PCA, SVM.*

---

<sup>1</sup> Facultad de Ingeniería. Universidad Militar Nueva Granada. Cra. 11 N° 101 -80. Bogotá, Colombia.  
E-mail: angela.mejia@unimilitar.edu.co; astrid.rubiano@unimilitar.edu.co

<sup>2</sup> Facultad de Ingeniería. Universidad Distrital Francisco José de Caldas. Cra. 7 N° 40-53. Bogotá, Colombia.  
E-mail: malzate@udistrital.edu.co

## INTRODUCCIÓN

El cáncer cervical se puede curar si se detecta y trata oportunamente [1]. La prueba de frotis de Papanicolaou (o prueba Pap) [2] intenta detectar este cáncer, o cambios neoplásicos previos, de acuerdo con propiedades como el color, la forma y la textura del núcleo y del citoplasma en células cervicales [3]. Sin embargo, la observación de estas características desde la imagen original es un proceso altamente subjetivo, ya que los resultados del diagnóstico pueden verse afectados por aspectos como el nivel de cansancio o estrés del médico [4]. Por esta razón, sería muy útil contar con una ayuda tecnológica que haga este tipo de mediciones de forma automática y precisa, y que ofrezca al médico criterios objetivos para reducir la naturaleza subjetiva del diagnóstico. Sin embargo, aún existen muchas dificultades por resolver en la construcción de dicha herramienta. En este artículo nos referiremos a dos de ellas.

Primero, para tomar estas mediciones de manera confiable es necesario poder segmentar las imágenes de células cervicales para distinguir el fondo, el citoplasma y el núcleo. Dicha segmentación permitiría medir de manera automática las características de interés que hayan demostrado ser discriminantes para la detección de estas lesiones. Sin embargo, a pesar de los muchos estudios que se han realizado, la segmentación de este tipo de imágenes sigue siendo un problema no resuelto debido a factores como la coloración inconsistente, el bajo contraste, la superposición de células, los elementos del fondo, etc. [5]. Segundo, aun contando con las imágenes correctamente segmentadas, existe una gran cantidad de variables que caracterizan aspectos relevantes de las células bajo estudio, como la forma del núcleo, la forma del citoplasma, sus tamaños relativos, las texturas de estos dos componentes e, incluso, la rugosidad de sus bordes. Se hace necesario, entonces, evaluar la pertinencia de estas variables como criterios para clasificar las células como normales o lesionadas [6].

Ya desde hace cerca de cuatro décadas se ha intentado resolver problemas relacionados con la segmentación y clasificación de células cervicales. Por ejemplo, en [7] se usaron modelos markovianos para representar la textura de los núcleos y concluyeron que es posible discriminar así núcleos normales y anormales si se seleccionan cuidadosamente los pasos

de cuantización; y en [8] se compararon diferentes métodos estándar de detección de bordes con un método que barría distintos niveles de comparación buscando el que mantuvieran más estable el perímetro de la célula. La principal conclusión de ambos trabajos era que la segmentación en imágenes de células cervicales requería mucha investigación debido a las múltiples dificultades por resolver. En octubre de 2014, 38 años después, se lanzó el segundo reto de segmentación de imágenes de citologías cervicales, ISBI 2015 [9], bajo la premisa de que segmentar células cervicales sobrelapadas es uno de los principales obstáculos por resolver en el análisis automático de imágenes microscópicas de frotis cervical. Por supuesto, en estas cuatro décadas se han logrado grandes avances. Entre los trabajos recientes se encuentra el mostrado en [5], por ejemplo, donde se usa una jerarquía multiescala de detección de bordes y, con base en mediciones de circularidad y homogeneidad, se determina si cada componente es célula o fondo. Posteriormente usan un clasificador binario para identificar el núcleo dentro de la célula. Aquí también los autores reconocen explícitamente que automatizar el proceso de segmentación de células cervicales sigue siendo un problema abierto debido a las complejidades de las estructuras en las células. En [10] se sigue insistiendo en la abrumadora dificultad en delimitar el núcleo dentro de la célula, debido a la amplia variabilidad de los núcleos entre células. Utilizan un procedimiento interactivo de determinación de regiones de interés sobre versiones de baja resolución y luego utilizan la máxima resolución para detectar las células dentro de esas regiones mediante filtros de mediana y detección de bordes. A las regiones detectadas se les aplican filtros morfológicos y se les mide la forma, el tamaño y la redondez. En [11] también se reconoce la segmentación de células cervicales como uno de los problemas de vieja data que no se han logrado resolver, por lo que recurren a la ayuda de un experto para poder solventar el amplio rango de variabilidad en las características de las imágenes. Usan filtros mediana y detector de bordes Canny, seguidos por una transformada Hough para detectar elipses y un algoritmo de conjuntos de nivel, reportando así resultados “promisorios”. En [12] y [13] los autores proponen un método de segmentación basado en una combinación de selección automática de dos niveles de comparación, seguida por el uso de contornos activos. En [14] proponen un proceso

de dos fases: en la primera fraccionan la imagen de la célula utilizando un algoritmo jerárquico y en la segunda determinan si cada fragmento pertenece al núcleo o al citoplasma. Aunque logran detectar el 100% de los núcleos, aún tienen dificultades segmentando los citoplasmas. En [15] proponen un algoritmo de segmentación basado en conjuntos de nivel para distinguir los núcleos y los citoplasmas. Sin embargo, los autores especifican que este algoritmo no debe ser considerado un método completo de segmentación, debido a las dificultades que aún persisten con la detección de los bordes del citoplasma en las células sobrepuestas. Como conclusión, el problema de segmentación de células en imágenes de frotis cervical sigue siendo un reto importante por resolver.

En este artículo presentamos un método de segmentación interactivo que permite clasificar el fondo, el citoplasma y el núcleo en imágenes de células aisladas. La interactividad consiste en que, con cada imagen a procesar, el usuario selecciona dos regiones que incluya a la célula aislada y al núcleo. Cada subimagen se separa en sus bandas espectrales (rojo, verde y azul) y a cada banda se aplica, de manera independiente, un filtro de mediana para eliminar ruidos y características irrelevantes, con ecualización del histograma antes y después del filtro. Al promediar la magnitud de los gradientes en cada banda se forma una imagen en la que se enfatizan los gradientes comunes en todas las bandas. Al reducir morfológicamente el grosor de las líneas así detectadas, los componentes conectados más grandes en cada subimagen suelen ser el citoplasma y el núcleo, respectivamente. Cuando no es así, se inicia un recorrido de las regiones segmentadas para reclasificarlas como núcleo, citoplasma o fondo de acuerdo con una medida de homogeneidad basada en el color y las distancias entre los centroides.

Luego de segmentada cada imagen se mide en ella un conjunto de 11 variables que representan el tamaño relativo del núcleo respecto del citoplasma, las formas de estos dos objetos, sus texturas y la rugosidad de sus bordes. Con estas 11 medidas se hace un proceso de extracción de características basado en análisis de componentes principales (PCA). Los parámetros resultantes se separan mediante un plano en el espacio de características, que se optimiza mediante una máquina de vectores

de soporte (SVM) para maximizar el margen (la distancia entre la frontera de decisión y el punto más cercano entre los datos disponibles) y así aumentar su capacidad de generalización [16]. Se considera que las células cuyos parámetros se ubiquen encima del plano presentan algún tipo de lesión neoplásica, mientras que las que se ubiquen debajo del plano son células normales.

Este detector se usa para analizar la relevancia de las variables medidas en las imágenes previamente segmentadas. Ante la dificultad de medir la información mutua entre las variables observadas y la clasificación de cada célula debido al bajo número de muestras, se usa como estimación la tasa de aciertos del detector cuando se le presentan células que no había visto durante el entrenamiento. Los cambios incrementales en el desempeño del clasificador se usan como estimadores de la información adicional que cada nueva variable añade respecto de las demás. Se concluye que las variables más relevantes para detectar lesiones neoplásicas en células cervicales se refieren a los tamaños relativos del núcleo y el citoplasma de las células. Las características que siguen en orden de relevancia se refieren a la forma de estos objetos, como su solidez o su excentricidad. Otras características que pueden ofrecer alguna información adicional son las características de textura y de rugosidad.

Para probar los métodos de segmentación, medición de características y clasificación de células, disponemos de 35 imágenes de células cervicales proporcionadas por el doctor Javier Rodríguez, del grupo de investigación Insight de la Universidad Militar y del Centro de Investigación de la Clínica del Country en Bogotá, Colombia. Estas imágenes han sido previamente clasificadas por el personal científico del grupo Insight, contando con 23 células normales y 12 células lesionadas.

En la siguiente sección describimos detalladamente el método de segmentación y presentamos sus resultados sobre la base de imágenes de que disponemos. En la tercera sección mostramos la utilidad de proceso de segmentación en la medición de diferentes características de las células. La cuarta sección se refiere al uso de un detector basado en PCA y SVM para analizar la relevancia de las medidas observadas en las imágenes segmentadas. En la última sección presentamos las conclusiones de nuestro trabajo.

## SEGMENTACIÓN DE IMÁGENES

Tratándose de imágenes de células cervicales, los métodos de segmentación global no son aplicables por la alta variabilidad entre imágenes [11]. Por esta razón conviene usar métodos locales, aunque problemas como la superposición de células también los puede afectar notoriamente [14]. Nuestros resultados experimentales señalan la dificultad de encontrar un método único que funcione igualmente bien en diferentes imágenes. Los problemas de iluminación o de contraste, por ejemplo, previenen la convergencia de métodos basados en la forma. Algunos otros métodos pueden funcionar bien en un gran número de células aisladas, pero fallan al encontrar células superpuestas [17]. Resulta irónico que estas grandes dificultades en la segmentación automática suelen no representar ningún problema para un ser humano, como muestra la Figura 1.

En esta sección presentamos el método interactivo de segmentación. La interactividad se da en un paso inicial en el que el usuario selecciona dos regiones rectangulares en la imagen donde se encuentran la célula de interés y su núcleo, respectivamente. Esta simple intervención elimina muchas de las fuentes de error en un proceso de segmentación completamente automático gracias a la simplicidad con que el ser humano resuelve problemas como el de la Figura 1. Luego se hace un proceso de segmentación morfológico que suele conducir a una segmentación correcta del núcleo y el citoplasma pero, si no es así, se entra a un proceso de refinamiento en el que las diferentes regiones resultantes de la segmentación original se van reclasificando como fondo, citoplasma o núcleo de acuerdo con una medida de homogeneidad basada en el color y las distancias entre los centroides. Es importante resaltar que la interactividad no se relaciona con



Figura 1. Un ser humano puede identificar aquí un cuadrado o una elipse inexistentes, engañosas para un proceso computacional.

la subjetividad inherente al proceso de diagnóstico de presencia de lesiones neoplásicas, pues se trata solo de determinar la ubicación de las células y sus núcleos, no de decidir la condición de la célula a partir de medidas tan difusas como texturas, formas, tamaños o bordes. Al final del proceso el sistema ofrecerá al médico medidas concretas de las variables de interés como apoyo tecnológico en el diagnóstico, reduciendo efectivamente la naturaleza subjetiva de dicho diagnóstico.

### Determinación interactiva de regiones de interés

La distinción entre fondo, citoplasma y núcleo podría realizarse con base en detección de bordes, para ello resultan útiles algunas operaciones básicas como la ecualización del histograma para enfatizar contrastes, los filtros pasa-altos convolucionales o morfológicos, etc. Sin embargo, en la experiencia obtenida con este trabajo estos métodos, aunque útiles, pocas veces llegan a segmentaciones apropiadas del citoplasma para un rango amplio de imágenes, sino que sus parámetros deben configurarse de manera particular para cada imagen. Por esta razón, en este trabajo se usa un primer proceso interactivo en el que el usuario indica un rectángulo en la imagen a procesar donde se encuentre una célula de interés y, dentro de él, otro rectángulo donde se encuentre el núcleo correspondiente. Este es un paso rápido y sencillo para el analista, a cambio del cual se elimina una de las principales fuentes de incertidumbre en el tratamiento de la imagen, pues una dificultad importante es la identificación automática de una célula aislada [12]. La Figura 2 muestra el resultado de este primer proceso de selección para una célula particular.

### Segmentación del núcleo

En casi todas las imágenes que se consideraron, el núcleo fue relativamente fácil de detectar dentro del rectángulo seleccionado por el usuario mediante técnicas estándar de detección de bordes. La Figura 3

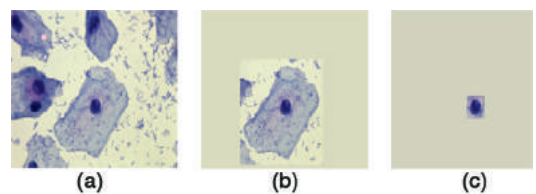


Figura 2. El usuario selecciona dos regiones rectangulares de interés en la imagen.



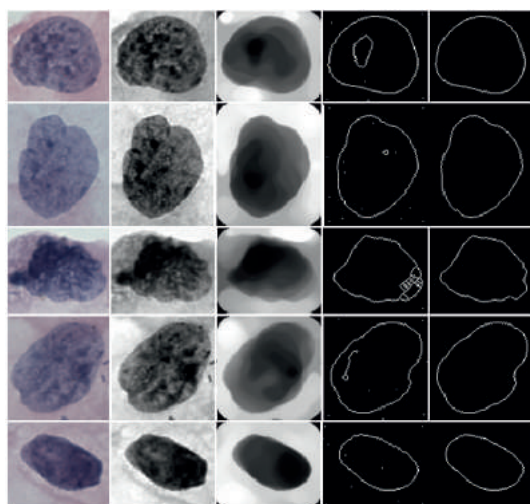


Figura 3. Proceso de segmentación del núcleo en cinco de las células.

muestra cinco etapas del proceso para cinco células particulares: A partir de la imagen original (primera columna) se construye la imagen en niveles de gris (segunda columna) y se aplica un filtro mediana (columna 3). La extracción de bordes mediante el método Canny (columna 4) suele retornar de una vez la región correspondiente al núcleo, pero en muchas ocasiones algunos aspectos de la estructura misma del núcleo o de objetos cercanos o superpuestos generan errores de clasificación, por lo que se seleccionó el componente conectado más grande (columna 5).

### Segmentación del citoplasma

La segmentación que permita distinguir el citoplasma del fondo de la imagen (o de otros citoplasmas superpuestos) no resulta fácil, pues los bordes entre citoplasmas o entre el citoplasma y el fondo suelen ser difusos. Además, la textura misma dentro del citoplasma puede generar bordes aún más intensos. Por eso, para la segmentación del citoplasma conviene aprovechar la información de color.

Cada imagen está compuesta por tres capas espectrales, rojo, verde y azul, con las cuales se construye una cuarta capa adicional de grises,  $\text{Gris} = 0.299 \times \text{Rojo} + 0.587 \times \text{Verde} + 0.114 \times \text{Azul}$ . Aunque entre casi todas ellas existe una alta correlación, en cada imagen hay alguna capa menos correlacionada, sugiriendo una pequeña información marginal. Esto es, usando la capa de grises se tiene un alto

porcentaje de toda la información sobre la imagen, pero existe, al menos, otra capa espectral que ofrece información adicional no contenida en la capa de grises: algunas capas en conjunto poseen más información que cada una de ellas por separado. Como esta información no redundante con la capa de grises puede estar en cualquiera de las otras tres capas (o, en algunos casos, hasta en dos de ellas), se procesan independientemente las cuatro capas a pesar de la alta redundancia.

La alta variabilidad de los valores de los píxeles en las regiones de fondo y citoplasma impide usar niveles de comparación estáticos para determinar a qué tipo de región pertenece cada pixel. Por esta razón se debe usar métodos adaptativos. Sin embargo, tanta es la variabilidad que ni siquiera un algoritmo de adaptación global resulta adecuado, por lo que proponemos un tipo de adaptación local. En un primer paso se usa el filtro mediana sobre las diferentes capas espectrales de la imagen (rojos, verdes, azules y la correspondiente versión de grises), con ecualización del histograma antes y después del filtro. De esta manera no solo se reducen varios tipos de ruido preservando la información de bordes, sino que se eliminan componentes irrelevantes de la imagen sin influir en la clasificación de las áreas en términos de fondo, citoplasma o núcleo. Sobre los componentes filtrados y ecualizados se calcula la magnitud del gradiente para notar cuánta información de interés hay en los bordes estimados en cada capa espectral. Aunque la información mutua entre ellas es muy alta, se alcanza a apreciar cierta información marginal en algunas capas. Así, en vez de aplicar la detección de bordes a cada capa espectral, se promedia la magnitud de los gradientes en cada capa y se forma una sola capa en la que se enfatizan los gradientes comunes en todas las capas. El promedio de las magnitudes de los gradientes de cada capa espectral muestra las regiones de la imagen en las que se debería buscar las fronteras entre las áreas de interés. Después de encoger morfológicamente el grosor de las líneas así detectadas y seleccionar el componente conectado más grande, en la mayoría de imágenes el citoplasma es la región encerrada por dicho componente. Sobre ella se superpone el área previamente seleccionada del núcleo. En muchas imágenes esta segmentación es aceptable, como muestra la Figura 4. Sin embargo, en otros casos se requiere el proceso adicional que describiremos a continuación.

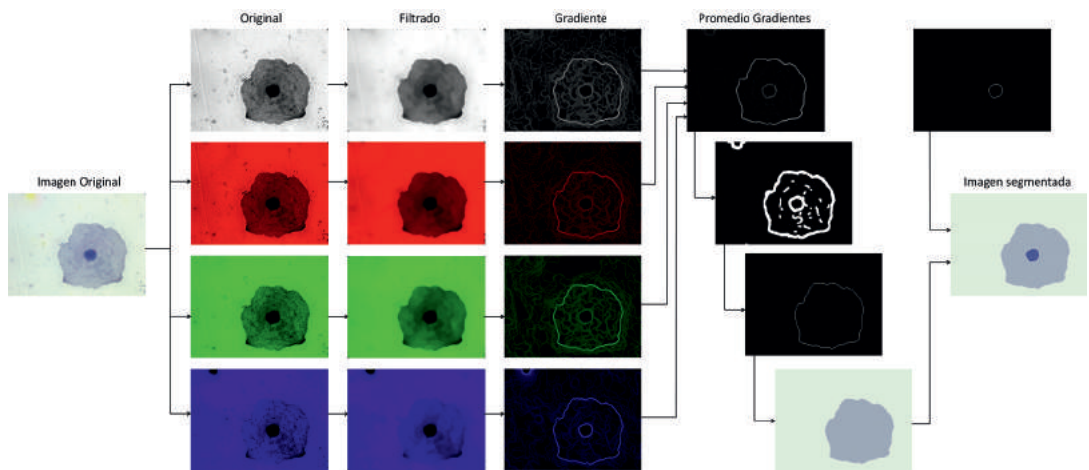


Figura 4. Proceso de segmentación del citoplasma y superposición del núcleo.

### Refinamiento de la segmentación

En algunos casos la segmentación anterior es insatisfactoria, como muestran las primeras dos imágenes de la Figura 5. En estos casos se decide hacer una segmentación con mayor resolución en la que el núcleo o el citoplasma resultan divididos en varias regiones. Posteriormente se recorren las regiones segmentadas y se reclasifican como núcleo, citoplasma o fondo de acuerdo con una medida de homogeneidad, como muestran las dos últimas imágenes de la Figura 5.

En efecto, al disponer de  $N$  regiones potencialmente significativas, podemos comparar el color de cada región con sus regiones vecinas y asociarlas de acuerdo con su similitud y la cercanía de sus centroides. Del conjunto de  $N(N-1)/2$  pares de segmentos, seleccionamos aquellos en que los pares sean vecinos adyacentes y, para cada par, calculamos la distancia entre los centroides y la diferencia de color. Entonces agrupamos el par más parecido de manera que se reduzca en uno el número de regiones. Este proceso se repite iterativamente con las  $N-1$  regiones restantes, iterando hasta obtener solo tres regiones, que conforman la segmentación deseada entre fondo, citoplasma y núcleo, como muestra la

última imagen de la Figura 5. La Figura 6 muestra nueve pasos intermedios de los 19 que se requirieron para pasar de una imagen dividida en 22 regiones a la célula correctamente segmentada (de la tercera a la cuarta imagen de la Figura 5).

La Figura 7 muestra el resultado de la segmentación para cada una de las 35 células en la base de datos de que disponemos.



Figura 6. Proceso de clasificación de múltiples regiones.



Figura 5. Segmentación en múltiples regiones.

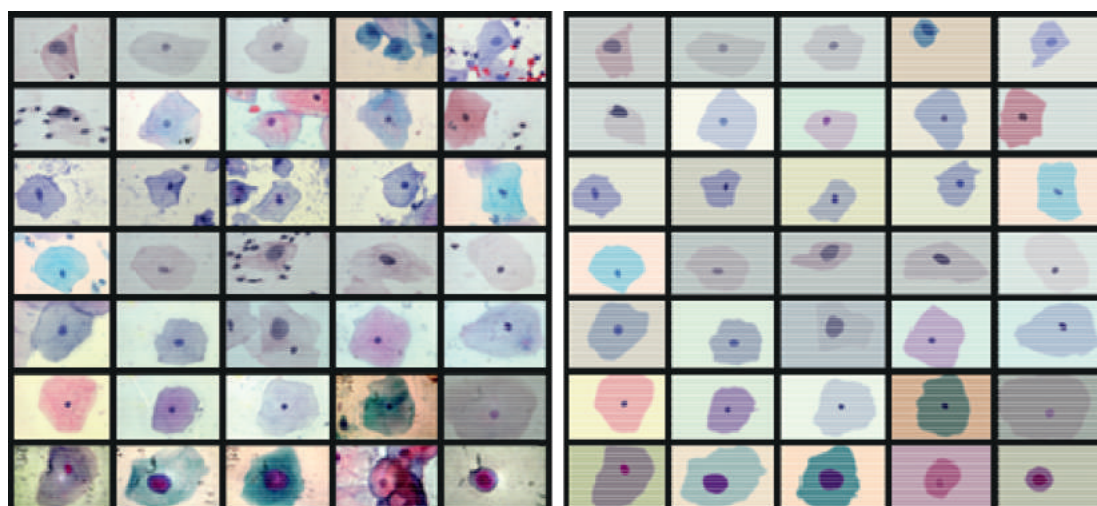


Figura 7. Las 35 células en la base de datos y sus respectivas segmentaciones.

Como aplicación de este proceso de segmentación, en la próxima sección se tomarán algunas medidas que resultan fáciles de adquirir en la imagen segmentada.

### MEDICIONES EN LA IMAGEN SEGMENTADA

Las características más sobresalientes en las imágenes segmentadas de la Figura 7 son los tamaños relativos entre el núcleo y el citoplasma, aunque también se nota alguna variabilidad en la forma, la rugosidad y la textura de cada región. La bondad del método de segmentación es que, una vez segmentadas las imágenes, resulta fácil medir de manera confiable características como las siguientes:

- $P_c$  = Perímetro del citoplasma, en píxeles
- $P_n$  = Perímetro del núcleo, en píxeles
- $A_c$  = Área del citoplasma, en píxeles
- $A_n$  = Área del núcleo, en píxeles
- $E_c$  = Excentricidad del citoplasma
- $E_n$  = Excentricidad del núcleo
- $S_c$  = Solidez del citoplasma
- $S_n$  = Solidez del núcleo
- $H_c$  = Entropía del citoplasma
- $H_n$  = Entropía del núcleo
- $d_c$  = Dimensión *box-counting* (*bc*) del borde del citoplasma
- $d_n$  = Dimensión *bc* del borde del núcleo
- $d_a$  = Dimensión *bc* de ambos bordes

El perímetro es el número de píxeles en el borde de la región de interés y el área incluye también los píxeles dentro de su interior. Por ejemplo, el objeto de la Figura 8 tiene un perímetro de 12 píxeles y un área de 24 píxeles.

La excentricidad se obtiene construyendo una elipse que tenga los mismos segundos momentos del objeto de interés y calculando la razón entre la distancia interfocal de la elipse y la longitud de su eje mayor [18]. Si la excentricidad es cero, se trata de un círculo. Si la excentricidad es 1, se trata de un segmento de línea recta. La Figura 9 muestra el cálculo de la excentricidad del citoplasma de una célula.

La solidez es la razón entre el área de la Figura y el área del casco convexo que la contiene [18]. Si la

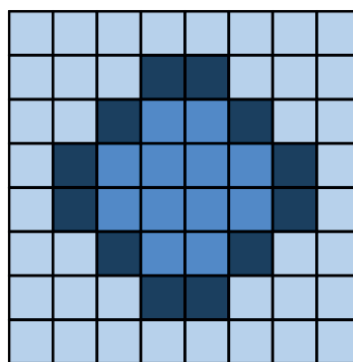


Figura 8. Medidas de longitud y área en píxeles. Perímetro =12, Superficie = 24.



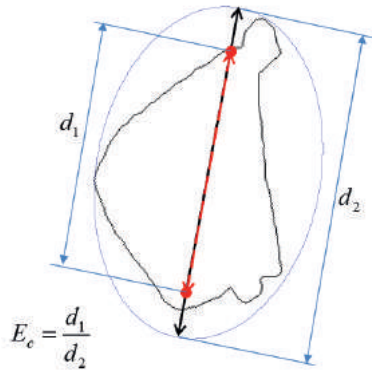


Figura 9. Cálculo de la excentricidad del citoplasma de una célula.

solidez es uno, se trata de un objeto convexo. Si la solidez es cero, el objeto está compuesto por curvas en el plano. La Figura 10 muestra el cálculo de la solidez del citoplasma de una célula.

La entropía es la información promedio de Shannon, calculada sobre la frecuencia relativa de los niveles de gris en cada región y normalizada respecto de ocho bits (la máxima entropía posible para 256 niveles de gris). Si la entropía es cero, el objeto completo está compuesto por un solo tono de gris. Si es uno, todos los tonos de gris se encuentran uniformemente distribuidos. Como muestra la Figura 11, la entropía es una medida indirecta de la textura del objeto.

Por último, la dimensión  $bc$  es la relación de potencia que existe entre la resolución de la imagen y la longitud del borde del objeto, en píxeles [13]. Si la imagen se representa con una resolución de  $1/\rho$  píxeles/milímetro, podríamos asociar un perímetro de  $P(\rho)$  píxeles con una longitud aproximada  $L(\rho)$

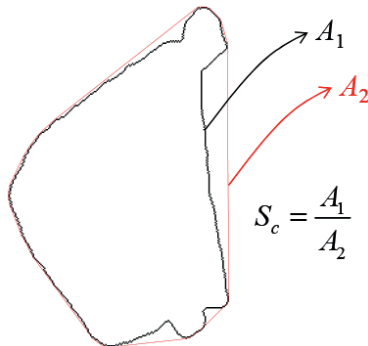


Figura 10. Cálculo de la solidez del citoplasma de una célula.

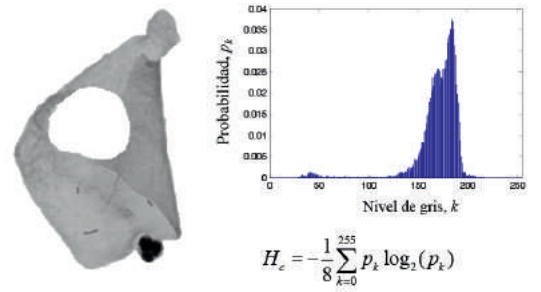


Figura 11. Cálculo de la entropía del citoplasma de una célula.

$= P(\rho) \times \rho^1$  milímetros, donde el exponente uno especifica la dimensión topológica del borde del objeto. Como muestra la Figura 12(a), esta medida es solo una aproximación, pero si la forma que se mide es suficientemente suave, bastaría con aumentar la resolución  $1/\rho$  (esto es, disminuir  $\rho$ ) para que la correspondiente medida de longitud sea cada vez más exacta, como se observa en la Figura 12(b). Sin embargo, si la curva a medir es rugosa, la longitud así definida puede crecer con la resolución en un rango apropiado de escalas, como muestra la Figura 12(c), de manera que la relación de potencia sería  $M(\rho) = P(\rho) \times \rho^d$ , donde  $M$  sería alguna “medida” en la “dimensión”  $d$ .

Una manera efectiva de detectar este fenómeno es por medio de la dimensión  $bc$  [19]: al tomar logaritmos en la relación anterior, se obtiene  $P(\rho) = \log(M) + d \log(1/\rho)$ , por lo que se pueden tomar muestras del número de píxeles  $P(\rho)$  que ocupa el borde para diferentes valores de  $\rho$ , y hacer una regresión lineal que permita estimar el valor de  $d$ . Aquí seguimos dicho procedimiento con versiones de las imágenes a  $(336 \times 448)$ ,  $(168 \times 224)$ ,  $(84 \times 112)$ ,  $(42 \times 56)$ ,  $(21 \times 28)$  y  $(10 \times 14)$  píxeles. Sin embargo, debido a la suavidad de los bordes mostrados en la Figura 7, las dimensiones obtenidas siempre fueron demasiado cercanas a 1 como para resultar útiles como criterio de detección de lesiones, a pesar de los resultados en [20]. Por eso preferimos escoger un rango pequeño de escalas donde se aprecie mejor la rugosidad, esto es, entre las seis escalas seleccionamos el grupo de tres escalas consecutivas que entreguen la mayor estimación de dimensión  $bc$ , pues la rugosidad de los bordes se da a nivel local en la escala (ver Figura 13). En [20] se procesan las medidas correspondientes para calcular un parámetro adicional, la armonía matemática

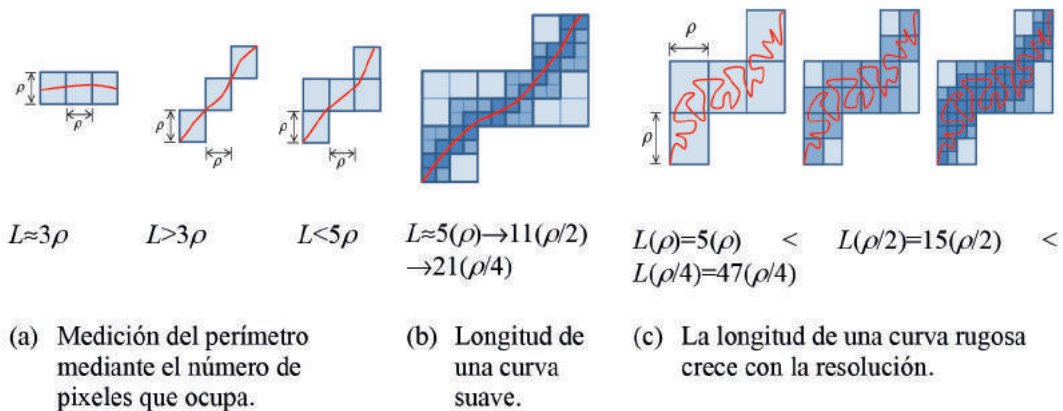


Figura 12. Relaciones entre el número de pixeles ocupado por una curva, su longitud y su rugosidad.

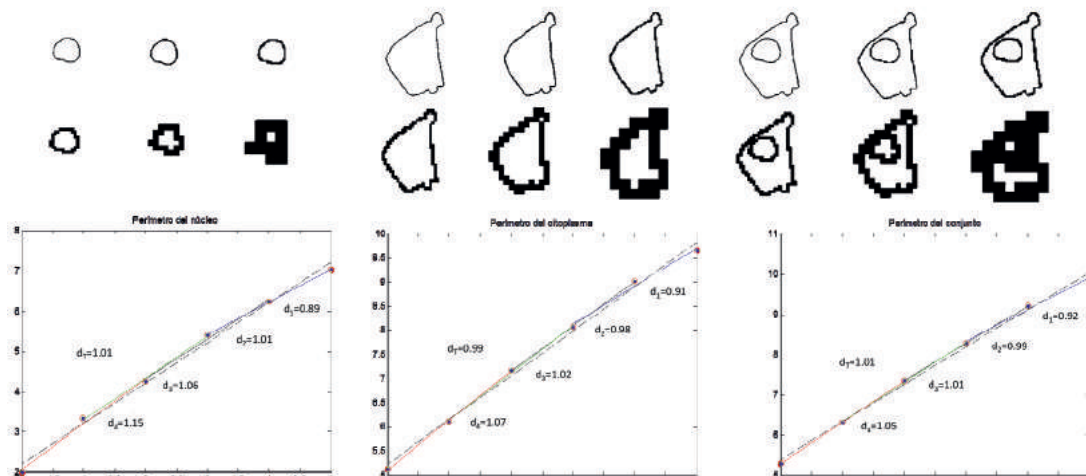


Figura 13. Cálculo de los parámetros de rugosidad mediante el algoritmo *box-counting*.

intrínseca (AMI), definida en [21] como el “grado de similitud o diferencia entre las unidades y las cifras significativas de las dimensiones fractales del núcleo, el citoplasma y el conjunto de ambos. Cuando la diferencia en las dimensiones fractales está en la unidad se denota con cero, cuando es en la primera cifra significativa con uno, en la segunda con dos y en la tercera con tres”. Aquí usaremos directamente las estimaciones de  $d_c$ ,  $d_n$  y  $d_a$ , con base en el principio básico de que el procesamiento de datos no incrementa la información [22], de manera que la capacidad discriminante de la AMI es menor o igual a la de los parámetros  $d_c$ ,  $d_n$  y  $d_a$  (además, los algoritmos *box-counting* no pueden ofrecer resultados precisos hasta el tercer decimal).

La solidez, la excentricidad, la entropía y la dimensión están convenientemente normalizadas

cerca a la unidad, pero no así el área ni la longitud. Como los tamaños de los objetos por sí mismos no son tan significativos como la relación entre ellos, hemos reducido las primeras cuatro medidas a solo dos medidas normalizadas, menos dependientes de las particularidades de la captura de las imágenes:

$$I_P = \text{Índice de perímetros} = (P_c - P_n) / (P_c + P_n)$$

$$I_A = \text{Índice de áreas} = (A_c - A_n) / (A_c + A_n)$$

En la siguiente sección construiremos un detector de lesiones con base en análisis de componentes principales (PCA) y máquinas de vectores de soporte (SVM), y utilizaremos las medidas de desempeño de este detector para discutir brevemente sobre la utilidad de las medidas anteriores en la detección de lesiones neoplásicas.

## ANÁLISIS DE LAS MEDICIONES EN LA IMAGEN SEGMENTADA

Al observar cómo se distribuyen las medidas en cada tipo de célula (Figura 14), se nota que algunas de ellas, como los índices de perímetros y áreas o la solidez del citoplasma, permiten distinguir bien unas células de otras al ocupar diferentes rangos de valores en cada caso. Otras medidas, como la solidez del núcleo o la dimensión del núcleo parecen prácticamente irrelevantes. Sin embargo, las medidas discriminantes pueden ser también redundantes, de manera que conjuntamente no ofrecen mucha más información que la que ofrece cada una de ellas por aparte, como se aprecia en el diagrama de dispersión de la Figura 15 para el caso de los índices de perímetro y área. En efecto, mediante una regresión cuadrática sería posible encontrar una estimación muy cercana de una medida en términos de la otra. La Figura 15 también muestra que bastaría comparar  $I_P$  con 0,654 para clasificar bien todas las células normales y errar en una célula lesionada, así como bastaría comparar  $I_A$  con 0,904 para clasificar bien todas las células lesionadas y errar en una célula normal. Sin embargo, no es posible trazar una línea recta en el plano  $I_P$ - $I_A$  que separe las células normales de las células lesionadas en el conjunto de las 35 células disponibles. ¿Podríamos

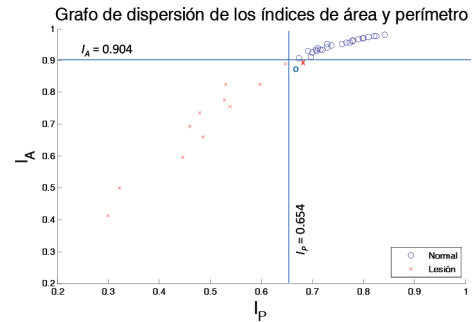


Figura 15. Los índices de perímetro y área son discriminantes, pero redundantes.

combinar linealmente las 11 variables para encontrar una única variable que permita detectar las células lesionadas mediante un único nivel de comparación?

Buscamos un vector  $\mathbf{u} = [u_1, u_2, \dots, u_{10}, u_{11}]^T \in \mathbb{R}^{11}$  tal que, al proyectar sobre él los  $N=35$  vectores  $\mathbf{x}(n) = [I_P(n), I_A(n), d_n(n), d_c(n), d_a(n), S_c(n), S_n(n), E_c(n), E_n(n), H_c(n), H_n(n)]^T \in \mathbb{R}^{11}$  para  $n=1, 2, \dots, N$ , obtengamos la máxima varianza entre las proyecciones. Suponiendo que el vector  $\mathbf{u}$  es un vector unitario ( $\mathbf{u}^T \mathbf{u} = \|\mathbf{u}\|^2 = 1$ ), la proyección del  $n$ -ésimo vector  $\mathbf{x}(n)$  (esto es, la proyección de la  $n$ -ésima célula) sobre  $\mathbf{u}$  es el producto punto  $\langle \mathbf{x}(n), \mathbf{u} \rangle = \mathbf{u}^T \mathbf{x}(n) = u_1 I_P(n) + u_2 I_A(n) + \dots + u_{10} H_c(n) + u_{11} H_n(n)$ .

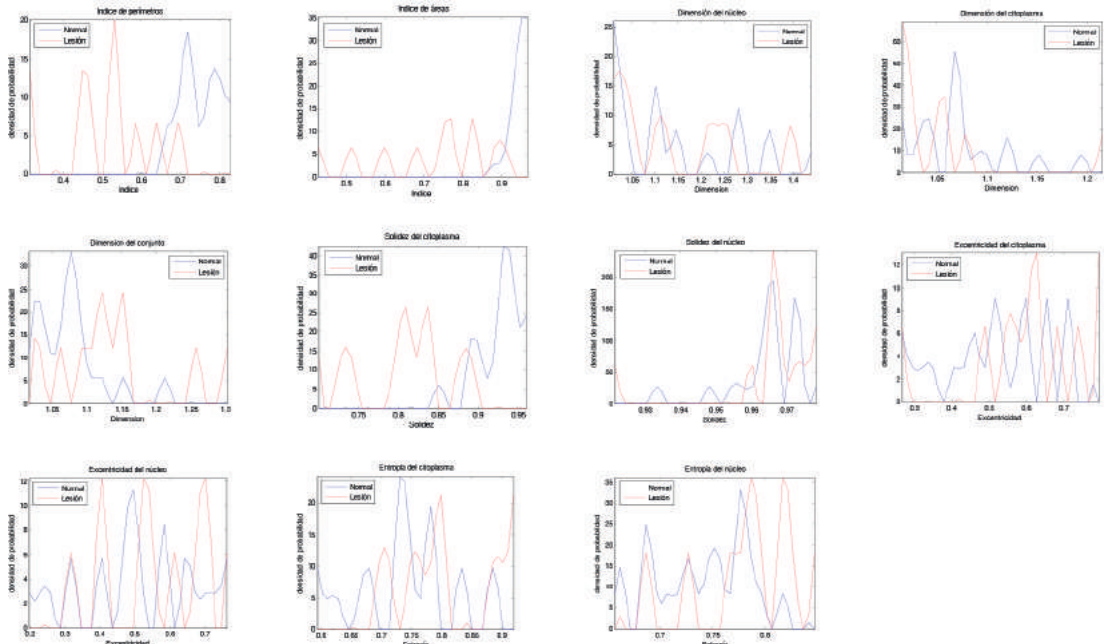


Figura 14. Funciones de densidad de probabilidad de las medidas observadas para cada tipo de célula.

$$\bar{x} = \frac{\sum_{n=1}^N [I_p(n), I_A(n), d_n(n), d_c(n), d_n(n), S_c(n), S_n(n), E_c(n), E_n(n), H_c(n), H_n(n)]^T}{N}$$

la media de las  $N$  proyecciones es  $\langle \bar{x}, \mathbf{u} \rangle = \mathbf{u}^T \bar{x}$ , de manera que la varianza de las proyecciones es

$$V = \frac{1}{n} \sum_{n=1}^N (\mathbf{u}^T x(n) - \mathbf{u}^T \bar{x})^2 = \frac{1}{n} \sum_{n=1}^N \mathbf{u}^T (x(n) - \bar{x})(x(n) - \bar{x})^T \mathbf{u} \\ = \mathbf{u}^T \left[ \frac{1}{n} \sum_{n=1}^N (x(n) - \bar{x})(x(n) - \bar{x})^T \right] \mathbf{u} = \mathbf{u}^T C \mathbf{u}$$

donde  $C = \frac{1}{N} \sum_{n=1}^N (x(n) - \bar{x})(x(n) - \bar{x})^T$  es la matriz

de covarianza de las 11 medidas originales. De esta manera el vector que estamos buscando es la solución del siguiente problema de optimización cuadrática:  $\max_{\mathbf{u} \in \mathbb{R}^{11}} \mathbf{u}^T C \mathbf{u}$  sujeto a  $\|\mathbf{u}\|^2 = 1$ . Al construir el lagrangiano,  $L(\mathbf{u}, \lambda) = \mathbf{u}^T C \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u})$ , podemos igualar a cero su derivada respecto de  $\mathbf{u}$  para obtener que el vector óptimo satisface  $C \mathbf{u} = \lambda \mathbf{u}$ , esto es,  $\mathbf{u}$  debe ser un vector propio de la matriz de covarianza. Al premultiplicar por  $\mathbf{u}^T$ , obtenemos que la varianza es  $\mathbf{u}^T C \mathbf{u} = \lambda \|\mathbf{u}\|^2 = \lambda$ , de manera que el  $\mathbf{u}$  óptimo es el vector propio unitario correspondiente al máximo valor propio de la matriz de covarianza. Al combinar linealmente las 11 medidas obtenidas con los elementos de este vector, se llega a una nueva variable,  $y_1$ , llamada “componente principal” [8]. Con esta variable las células normales se pueden separar perfectamente de las células lesionadas en la base de imágenes: la célula lesionada de máximo valor se ubica en  $y_1 = 0,655$  y la célula normal de mínimo valor se ubica en  $y_1 = 0,715$ , de manera que conviene comparar  $y_1$  con 0,685 para clasificar una célula como normal (si  $y_1$  es mayor) o como lesionada (si  $y_1$  es menor), maximizando el margen entre ellas, como muestra la Figura 16.

Para aumentar el margen podemos añadir otros componentes cuya información sea adicional a la proporcionada por  $y_1$ . Porque los 11 vectores propios

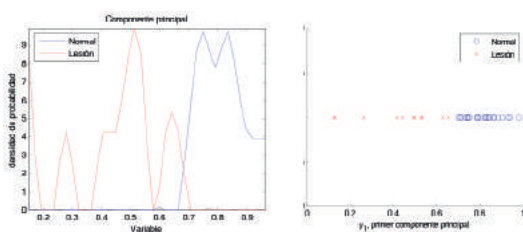


Figura 16. Clasificación de células con un componente principal.

de la matriz de covarianza forman una base ortogonal para el espacio vectorial expandido por las columnas de la matriz de covarianza, cada nuevo vector propio, por orden decreciente de valor propio, representa los coeficientes de la combinación convexa de medidas que aporta una nueva información no contenida en las anteriores combinaciones (o, al menos, no correlacionada linealmente con las anteriores). Así, el segundo componente principal se obtiene mediante la combinación convexa de las 11 medidas originales con los coeficientes tomados de la segunda columna de la matriz de covarianza (la correspondiente al segundo valor propio en magnitud). Cuando se usa este segundo componente aumenta el margen y, por tanto, la capacidad de generalización del detector, como muestra la Figura 17. Con tres componentes principales el problema de separar las células con una frontera de decisión que maximice el margen conduce a la solución de la Figura 18. Efectivamente, en el espacio de los tres componentes principales las células de la base de imágenes con que contamos se separan fácilmente mediante un plano, aumentando el margen y la capacidad de generalización.

Es interesante notar que en estos tres componentes principales las variables de mayor peso son los índices de perímetros y áreas, las excentricidades de núcleo y citoplasma y la dimensión  $bc$  del citoplasma. Si comparamos con la Figura 14, notamos que solo los índices de perímetros y áreas son realmente discriminantes por sí mismos, mientras que las excentricidades y las dimensiones  $bc$  no parecen serlo. En cambio, la solidez del citoplasma pierde relevancia en presencia de las otras variables, a pesar de que basta compararla con 0,885 para clasificar

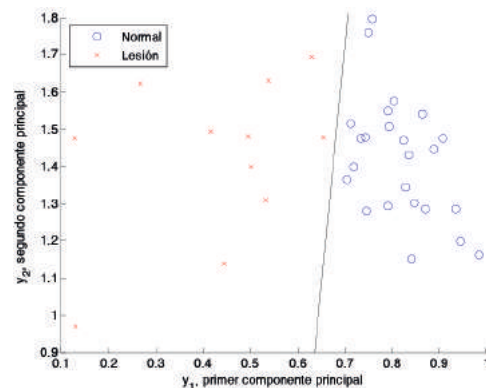


Figura 17. Clasificación de células con dos componentes principales.



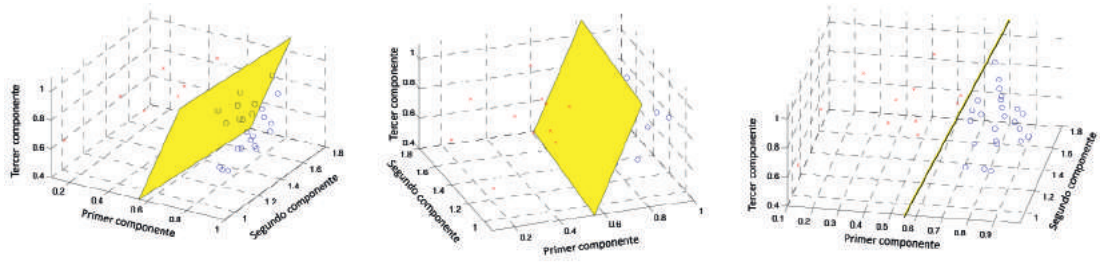


Figura 18. Clasificación de células con tres componentes principales.

correctamente todas las células con excepción de una normal y una lesionada.

En consecuencia, se hace necesario analizar cada variable tanto aislada como en conjunto con otras variables. Para hacer este análisis se usa el clasificador recién descrito; esto es, un proceso de análisis de componentes principales selecciona las tres mejores combinaciones convexas de las variables de entrada, y esas tres combinaciones se usan como entrada de una máquina de vectores de soporte que separe óptimamente las células normales y las células lesionadas. Sin embargo, si se le muestran al clasificador todas las células, él aprende a clasificarlas sin dificultad con el 100% de aciertos. Por eso se le mostraron solo seis células lesionadas y 12 células normales debidamente clasificadas (datos de entrenamiento) y se midió el desempeño con el número de aciertos en las 17 células restantes (datos de prueba). Se escogieron al azar 2000 configuraciones de los datos de entrenamiento y se promediaron las 2000 tasas de acierto en los correspondientes datos de prueba. Esta medición se repitió para cada una de las 2047 combinaciones de variables de entrada (11 variables individuales, 55 pares de variables, 165 tríos, 330 cuartetos, etc.).

La Figura 19 muestra las variables individuales ordenadas de la menos relevante a la más relevante, de acuerdo con el promedio de aciertos en los 2000 experimentos. Los resultados concuerdan perfectamente con la interpretación que hacíamos de la Figura 14. Sin embargo, al considerar conjuntos de variables se nota que la información marginal que ofrece cada variable puede estar ya contenida en otras. La Tabla 1 enumera las mejores combinaciones de una, dos, tres y hasta once variables, respecto de la tasa de aciertos que alcanzan conjuntamente cuando se usan como entrada al clasificador. Las

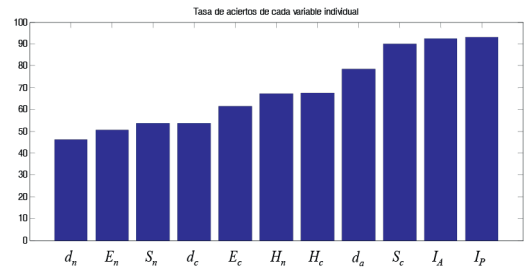


Figura 19. Desempeño del detector basado en una variable aislada.

tasas de aciertos de todas las 2047 combinaciones de variables de entrada se presentan en la Figura 20.

Por ejemplo, aunque las dos variables que individualmente ofrecen más información sobre la clasificación de las células son  $I_P$  e  $I_A$ , la mejor combinación de dos variables es  $(I_P, S_C)$ , no  $(I_P, I_A)$ . Esto implica que, aunque  $I_A$  revela más información acerca del tipo de célula que  $S_C$ , al condicionar en  $I_P$ , es mayor la información novedosa (no contenida ya en  $I_P$ ) proporcionada por  $S_C$  que por  $I_A$ .

Si  $C$  es la variable de Bernoulli que vale 1 para una célula lesionada y cero para una célula normal, diríamos, desde el punto de vista de la teoría de la información, que  $I(C; I_P) > I(C; I_A) > I(C; S_C)$  pero  $I(C; I_P, S_C) > I(C; I_P, I_A)$ , indicando que  $I(C; S_C | I_P) > I(C; I_A | I_P)$ :  $I_A$  tiene mayor información sobre  $C$  que  $S_C$ , pero la información proporcionada por  $S_C$  que no se encuentra en  $I_P$  es mayor que la información proporcionada por  $I_A$  que no se encuentra en  $I_P$ . No pretendemos que la tasa de aciertos del conjunto de variables  $x$  se pueda considerar una medida de la incertidumbre sobre el tipo de una célula,  $C$ , revelada por  $x$ , aunque deben ser cantidades íntimamente relacionadas. Si tuviésemos un clasificador ideal capaz de extraer toda la información relativa a  $C$  contenida en las variables de entrada, la información mutua y



Tabla 1. Conjuntos de  $n$  variables más discriminantes,  $n \in \{1, 2, 3, \dots, 11\}$ .

Número variables	Mejor conjunto de variables	Tasa de aciertos
1	$I_P$	92.9081
2	$I_P, S_c$	94.3176
3	$I_P, S_c, H_c$	94.3529
4	$I_A, S_c, E_c, E_n$	95.1691
5	$I_A, d_n, S_c, E_c, E_n$	95.5294
6	$I_A, d_c, d_a, S_c, E_c, E_n$	96.0647
7	$I_A, d_c, d_a, S_c, S_n, E_c, E_n$	96.1434
8	$I_A, d_c, d_a, S_c, S_n, E_c, E_n, H_n$	96.0706
9	$I_P, d_n, d_c, d_a, S_c, S_n, E_c, E_n, H_n$	95.6294
10	$I_P, d_n, d_c, d_a, S_c, S_n, E_c, E_n, H_c, H_n$	94.3588
11	$I_P, I_A, d_n, d_c, d_a, S_c, S_n, E_c, E_n, H_c, H_n$	93.1165

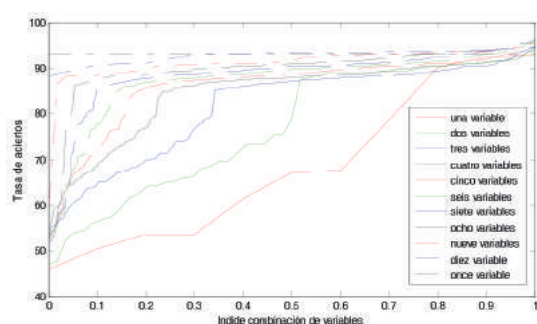


Figura 20. Tasas de aciertos de todos los posibles conjuntos de variables de entrada (el eje horizontal es un índice normalizado que representa las posibles combinaciones de las variables).

la tasa de aciertos serían dos cantidades directamente proporcionales; pero nuestro clasificador no puede extraer toda la información (solo usa combinaciones lineales) y, en cambio, sí pierde información durante el proceso de análisis de componentes principales, por ejemplo. Aun así, la Figura 20 y la Tabla 1 muestran que el clasificador cumple un excelente papel extrayendo casi toda la información mutua entre el tipo de célula y las variables de entrada.

Si queremos añadir una variable más, no deberían ser las siguientes dos variables más discriminantes ( $I_A$  y  $d_a$ ), sino  $H_c$ . Del conjunto de las tres mejores variables, ( $I_P, S_c, H_c$ ), al conjunto de las cuatro mejores, ( $I_A, S_c, E_c, E_n$ ), no bastó con añadir una cuarta variable sino que se debió cambiar dos variables tan importantes como  $I_P$  y  $H_c$  por tres variables más débiles individualmente,  $I_A, E_c$  y  $E_n$ . Se diría que  $I_P$  y  $H_c$  tienen más información sobre  $C$  adicional a la

proporcionada por  $S_c$  que cualquier par de variables en  $\{I_A, E_c, E_n\}$ , pero  $I_A, E_c$  y  $E_n$  tienen más información sobre  $C$  adicional a la proporcionada por  $S_c$  que cualquier trío de variables en  $\{I_P, H_c, I_A, E_c, E_n\}$ . Esto es, aunque  $I(C; I_P, H_c | S_c) > \max[I(C; I_A, S_c | S_c), I(C; I_A, E_n | S_c), I(C; E_c, E_n | S_c)]$ , también es cierto que  $I(C; I_A, E_c, E_n | S_c) > \max[I(C; I_P, H_c, I_A | S_c), I(C; I_P, H_c, E_c | S_c), I(C; I_P, H_c, E_n | S_c)]$ .

Por último, las variables observadas se pueden agrupar según el tipo de características que miden en la célula: variables de tamaño ( $I_P$  e  $I_A$ ), de textura ( $H_c$  y  $H_n$ ), de forma ( $S_c, S_n, E_c, E_n$ ) y de dimensión ( $d_c, d_n, d_a$ ). La Tabla 2 presenta las tasas de aciertos, falsos positivos y falsos negativos para la mejor combinación de variables que incluyan al menos una de cada tipo. Las características más relevantes son las de tamaño y forma que, de hecho, se combinan en el mejor grupo de cuatro variables de entrada,  $I_A, S_c, E_c$  y  $E_n$ , con una tasa de aciertos de 95,17%, y unas bajas tasas de falsos positivos (células normales clasificadas como lesionadas) y falsos negativos (células lesionadas clasificadas como normales). Las variables menos relevantes son las de textura —aunque debe considerarse que estamos midiendo la textura mediante una variable que habla solo de la cantidad de niveles de gris, no de la correlación entre ellos—. Al añadir los parámetros de rugosidad a los de forma y tamaño, la tasa de aciertos aumenta en 0,97%, mejorando también la tasa de falsos positivos y falsos negativos. Esto implica que, aunque los parámetros de rugosidad no son muy relevantes por sí mismos, añaden una pequeña información adicional que no había sido capturada por los parámetros de tamaño y forma. Los parámetros de textura solo mejoran la tasa de aciertos en 0,55%, indicando

Tabla 2. Características más relevantes de las células para detección de lesiones.

Tipos de variables	Combinación óptima	Aciertos	Falsos positivos	Falsos negativos
Tamaño	$I_p$	92.91%	4.9%	12.35%
Dimensión	$d_c$ y $d_a$	78.75%	16.53%	32.58%
Forma	$S_c$	90.08%	7.48%	15.76%
Textura	$H_c$ y $H_n$	69.7%	29.13%	33.1%
Tamaño y dimensión	$I_p$ y $d_n$	91.46%	5.81%	15.1%
Tamaño y forma	$I_A$ , $S_c$ , $E_c$ y $E_n$	95.17%	4.31%	6.08%
Tamaño y textura	$I_A$ y $H_n$	91.32%	7.9%	10.56%
Dimensión y forma	$d_c$ y $S_c$	88.36%	7.48%	21.64%
Dimensión y textura	$d_c$ , $d_a$ y $H_c$	79.75%	15.88%	30.74%
Forma y textura	$S_c$ y $H_c$	89.46%	5.94%	21.58%
Tamaño, dimensión y forma	$I_A$ , $d_c$ , $d_a$ , $S_c$ , $S_n$ , $E_c$ y $E_n$	96.14%	2.98%	5.96%
Tamaño, dimensión y textura	$I_p$ , $d_a$ y $H_n$	88.75%	6.44%	22.8%
Tamaño, forma y textura	$I_A$ , $S_c$ , $S_n$ , $E_c$ , $E_n$ y $H_n$	95.72%	3.68%	5.7%
Dimensión, forma y textura	$d_n$ , $d_c$ , $S_c$ , $E_c$ y $H_n$	89.32%	7.48%	18.36%
Tamaño, dimensión, forma y textura	$I_A$ , $d_c$ , $d_a$ , $S_c$ , $E_c$ , $E_n$ y $H_n$	96.14%	3.22%	5.41%

que su información adicional es menor que la de los parámetros de rugosidad.

## RESULTADOS

1. En vez de pasar por largos procesos genéricos que lentifican la segmentación, una breve interacción con el usuario para identificar las regiones donde se encuentran la célula y el núcleo permite eliminar una de las principales fuentes de error en la segmentación.
2. Sobre la imagen marcada por el usuario es posible recorrer los componentes obtenidos por un sencillo algoritmo de detección de bordes para clasificarlos como núcleo, citoplasma o fondo de acuerdo con una medida de homogeneidad basada en la adyacencia, en el color y en la distancia entre centroides de cada componente. La precisión en la segmentación hace más confiables las medidas obtenidas.
3. Aunque una adecuada combinación lineal de las 11 medidas obtenidas es suficiente para clasificar adecuadamente las 35 células disponibles en una dimensión, es posible aumentar el margen (la capacidad de generalización) usando medidas multidimensionales basadas en diferentes combinaciones lineales.
4. Con solo 35 muestras no es posible estimar la información mutua que las distintas medidas ofrecen acerca de la clasificación de las células. Sin embargo, agrupando de manera aleatoria grupos de 18 células, 6 lesionadas y 12 sanas, se puede construir cerca de mil doscientos millones de combinaciones de 18 datos de entrenamiento y 17 datos de prueba. Midiendo el desempeño promedio de diferentes SVM, podemos hacer una estimación de la cantidad de información que diferentes grupos de variables tienen sobre la clasificación de las células. La estimación es solo indirecta, pues la tasa de aciertos no es monótonamente creciente con el número de variables.
5. De acuerdo con esa medida indirecta, las características más relevantes son los índices de perímetro y área, aunque ellos son ampliamente redundantes, pues la información conjunta del índice de perímetro y la solidez del citoplasma es mayor que la de ambos índices en conjunto.
6. La mayor información conjunta se encuentra en la característica de tamaño  $I_A$ , las características de forma  $S_c$ ,  $S_n$ ,  $E_c$  y  $E_n$ , y en las de dimensión,  $d_c$  y  $d_a$ . Las características de textura,  $H_c$  y  $H_n$ , son prácticamente irrelevantes.

## CONCLUSIONES

En este artículo se presentó una técnica de segmentación interactiva de imágenes de células cervicales para distinguir fondo, citoplasma y núcleo, se describió un detector de lesiones neoplásicas en este tipo de células, y se analizó la relevancia de distintos grupos de variables para esta aplicación particular.

Aunque el problema de la segmentación ha eludido una solución definitiva desde el procesamiento digital de imágenes, con una breve participación del usuario logramos segmentar satisfactoriamente las 35 células de la base de datos de que dispusimos. Otro aporte es el uso de las distintas bandas espectrales de la imagen, las cuales complementan la información fundamental proporcionada por la escala de grises.

Un segundo resultado tiene que ver con la medición de características a partir de la versión segmentada de la imagen. La medida característica más discriminante es el tamaño relativo del núcleo respecto del citoplasma, para ello se compararon los perímetros y las superficies de los objetos de interés, obteniendo dos parámetros altamente discriminantes pero redundantes. Por otro lado, ante reportes de evidencias de que algunas medidas de fractalidad del núcleo y el citoplasma podrían tener información discriminante, se hizo el cálculo automático de las dimensiones  $bc$  del perímetro de los núcleos y los citoplasmas en el estrecho rango de escalas en las que se percibe alguna rugosidad. Este “parámetro de rugosidad” presenta una mayor varianza entre células que la dimensión  $bc$ . Se consideraron también medidas de forma como la solidez y la excentricidad de los núcleos y los citoplasmas, y medidas relacionadas con la textura, como las entropías de los niveles de gris. Para probar la relevancia de estas variables en la detección de lesiones neoplásicas se construyó un detector basado en análisis de componentes principales (PCA) y máquinas de vectores de soporte (SVM). Usando el desempeño del detector como una medida indirecta de estimación de la información mutua entre las variables de entrada y la condición de las células se concluyó que las medidas más relevantes son los parámetros de forma y tamaño, mientras que los parámetros de rugosidad y textura también incluyen alguna información adicional, pero reducida.

## AGRADECIMIENTOS

Este trabajo fue financiado por la Universidad Militar Nueva Granada mediante el proyecto de investigación ING1070. Agradecemos al grupo de investigación Insight por la contribución en el desarrollo de este proyecto de investigación y en particular al líder del grupo, el Dr. Javier Rodríguez.

## REFERENCIAS

- [1] American Cancer Society. “Cáncer de cuello uterino: detección temprana y prevención”. 7 de mayo de 2013. Fecha de consulta: 15 de septiembre de 2013. URL: <http://www.cancer.org/acs/groups/cid/documents/webcontent/002580-pdf.pdf>
- [2] G.N. Papanicolaou. “The diagnostic value of vaginal smears in carcinoma of the uterus”. American Journal of Obstetrics and Gynecology. Vol. 42, pp. 193-205. 1941.
- [3] D. Saslow. “American Cancer Society, American Society for Colposcopy and Cervical Pathology and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer”. A Cancer Journal for Clinicians. Vol. 62, Issue 3, pp. 147-172. 2012.
- [4] M. Peluffo. “Revoluciones Conceptuales en Patología Cervical”. Ascune Hermanos. 2a edición. Buenos Aires, Argentina. Noviembre de 2003.
- [5] A. Genstav, S. Aksoy and S. Onder. “Unsupervised segmentation and classification of cervical cell images”. Pattern recognition. Vol. 45, pp. 4151-4168. 2012.
- [6] A. Chaddad, C. Tanougast, A. Golato and A. Dandache. “Carcinoma cell identification via optical microscopy and shape feature analysis”. Biomedical Science and Engineering. Vol. 6, pp. 1029-1033. November, 2013.
- [7] Norman J. Pressman. “Markovian Analysis of Cervical Cell Images”. Journal of Histochemistry and Cytochemistry. Vol. 24, Issue 1, pp. 138-144. 1976.
- [8] R.L. Cahn, R.S. Poulsen and G. Toussaint. “Segmentation of cervical cell images”. J. Histochem Cytochem. Vol. 25, Issue 7, pp. 681-8. July, 1977.

- [9] Australian Research Council. "The Second Overlapping Cervical Cytology Image Segmentation Challenge - ISBI 2015". Date of visit: December 15, 2014. URL: [http://cs.adelaide.edu.au/~zhi/isbi15\\_challenge/](http://cs.adelaide.edu.au/~zhi/isbi15_challenge/)
- [10] C. Bergmeir, M. García Silvente, J. Esquivias López-Cuervo and J.M. Benítez. "Segmentation of cervical cell images using mean-shift filtering and morphological operators". Proc. SPIE 7623, Medical Imaging 2010: Image Processing, 76234C. March 12, 2010.
- [11] C. Bergmeir, M. García and J. Benítez, "Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework". Computer Methods and Programs in Biomedicine. Vol. 107, pp. 497-512. September, 2012.
- [12] J. Fan, S. Li and C. Zhang "Color Cell Image Segmentation Based on Chan-Vese Model for Vector-Valued Images". Journal of Software Engineering and Applications. Vol. 6, pp. 554-558. October, 2013.
- [13] J. Fan, R. Wang, S. Li and C. Zhang. "Automated Cervical Cell Image Segmentation using Level Set Based Active Contour Model". 12th International Conference on Control, Automation, Robotics & Vision, Guangzhou, China. December 5-7, 2012.
- [14] A. Kale and S. Aksoy. "Segmentation of Cervical Cell Images". 20th IAPR International Conference on Pattern Recognition. Istanbul, Turkey. August 23-26, 2010. URL: <http://dx.doi.org/10.1109/ICPR.2010.587>
- [15] Z. Lu, G. Carneiro and A. Bradley. "Automated nucleus and cytoplasm segmentation of overlapping cervical cells". Medical Image computing and computer-assisted intervention, Springer Verlag, Lecture notes on computer science. Vol. 8149, pp. 452-460. 2013.
- [16] C.M. Bishop. "Pattern recognition and machine learning". Springer Verlag. 2006.
- [17] J. Jantzen and G. Dounias. "Analysis of Pap-smear Image Data". Proceedings of the Nature-Inspired Smart Information Systems, NiSIS. Tenerife, Spain. 2006.
- [18] M.E. Plissiti, C. Nikou and A. Charchanti. "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images". Pattern Recognition Letters. Vol. 32, Issue 6, pp. 838-853. April, 2011.
- [19] H.-O. Peitgen, H. Jürgens and D. Saupe. "Chaos and Fractals: new frontiers of science". Springer Verlag, Second edition. New York, United States. 2004.
- [20] J.O. Rodríguez. "Nuevo método de ayuda diagnóstica con geometría fractal para células preneoplásicas del epitelio escamoso cervical". Revista UDCA. Vol. 14 N° 1, pp. 15-22. Junio de 2011.
- [21] J. Rodríguez-Velásquez, E. Prieto, L. Ortiz, M. Ronderos y C. Correa, "Diagnóstico matemático de ecocardiografías pediátricas con medidas de dimensión fractal evaluadas con armonía matemática intrínseca". Revista Colombiana de Cardiología. Vol. 17 N° 2, pp. 79-86. Marzo-Abril de 2010.
- [22] J. Thomas and T. Cover. "Elements of Information Theory". John Wiley and Sons. 2nd edition. New York, United States. 2006.