

**HERRAMIENTA BASADA EN MÁQUINAS DE SOPORTE VECTORIAL PARA
LA CLASIFICACIÓN DE CÉLULAS ESCAMOSAS DE CITOLOGÍAS
CÉRVICO UTERINAS**

**LAURA YADIRA VALBUENA RODRÍGUEZ
MILTON ANTONIO CAVIEDES SANABRIA**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2009**

**HERRAMIENTA BASADA EN MÁQUINAS DE SOPORTE VECTORIAL PARA
LA CLASIFICACIÓN DE CÉLULAS ESCAMOSAS DE CITOLOGÍAS
CÉRVICO UTERINAS**

**LAURA YADIRA VALBUENA RODRÍGUEZ
MILTON ANTONIO CAVIEDES SANABRIA**

**Trabajo de Grado para Optar por el Título de
Ingeniero(a) de Sistemas**

**Director
Víctor Eduardo Martínez Abaunza
Magíster en Ingeniería: Área Informática y Ciencias
De la Computación**

**Codirectora
Olga Mercedes Álvarez Ojeda
MD. Especialista en Patología**

**UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2009**

A mi familia, gracias por el apoyo incondicional
A Germán por ser más que un amigo

Laura Yadira Valbuena Rodríguez

A mi mamá y a mi esposa

Milton Antonio Caviedes Sanabria

AGRADECIMIENTOS

A Víctor Eduardo Martínez Abaunza por su ayuda y paciencia.

A la Doctora Olga Mercedes Álvarez Ojeda por gran acompañamiento.

A Germán Andrés Morales España por su aporte con la herramienta de clasificación.

CONTENIDO

| | |
|---|------------|
| 1. INTRODUCCIÓN | 1. |
| 2. PRESENTACIÓN DEL PROYECTO | 2. |
| 2.1 OBJETIVOS | 2. |
| 2.1.1. Objetivo general | 2. |
| 2.1.2. Objetivos específicos | 2. |
| 2.2. JUSTIFICACIÓN | 3. |
| 3. MARCO TEÓRICO | 4. |
| 3.1. CITOLOGÍA CÉRVICO UTERINA | 4. |
| 3.1.1. Coloración de Papanicolaou | 4. |
| 3.1.2. Hematoxilina Biopur: Colorante nuclear. | 5. |
| 3.2. COMPONENTES CELULARES EN LA CITOLOGÍA CÉRVICO UTERINA | 5. |
| 3.2.1. Células Cervicales | 5. |
| 3.2.1.1. Células Superficiales | 6. |
| 3.2.1.2. Células Intermedias | 7. |
| 3.2.1.3. Células Para-Basales | 8. |
| 3.3. ESTRUCTURAS CELULARES QUE SON PARÁMETROS DE EVALUACIÓN .. | 9 . |
| 3.4. PROCESOS PATOLÓGICOS QUE AFECTAN LAS CÉLULAS | 10. |
| 3.4.1. Efectos morfológicos sobre células que resultan en una actividad normal. Euplasia | 10. |
| 3.4.2. Efectos morfológicos sobre células de procesos asociados con cáncer. | 10. |
| 3.4.3. Efectos morfológicos sobre células de procesos retrogresivos resultando en una disminución de la actividad. Retroplasia | 11. |
| 3.4.4. Efectos morfológicos sobre células de procesos progresivos resultando en actividad aumentada. Proplasia. | 11. |
| 3.4.5. Efectos morfológicos sobre células de procesos cambiantes y mixtos. | 11. |
| 3.5. DIAGNÓSTICO EN CITOLOGÍA | 11. |
| 3.5.1. Sistema Bethesda 2001 | 13. |
| 3.6. TRATAMIENTO DIGITAL DE IMÁGENES | 14. |
| 3.6.1. Fundamentos | 14. |
| 3.6.2. Modelos de Color | 14. |

| | |
|---|------------|
| 3.6.3. Modelo de color RGB..... | 15. |
| 3.6.4. Modelo de color HSL..... | 16. |
| 3.6.5. Conversión del Modelo RGB a HSL..... | 17. |
| 3.7. ETAPAS DEL TRATAMIENTO DIGITAL DE IMÁGENES..... | 19. |
| 3.7.1. Adquisición | 20. |
| 3.7.2. Pre-procesamiento..... | 20. |
| 3.7.3. Segmentación | 21. |
| 3.7.4. Descripción | 25. |
| 3.7.5. Interpretación y Reconocimiento | 29. |
| 3.8. MÁQUINAS DE SOPORTE VECTORIAL (SVM)..... | 30. |
| 3.8.1. Metodología de las SVM..... | 32. |
| 3.8.2. Análisis del caso linealmente separable..... | 33. |
| 3.8.2.1 Hiperplano clasificador óptimo | 33. |
| 3.8.2.2 Solución al problema de optimización con restricciones..... | 35. |
| 3.8.3. Análisis del caso linealmente no separable..... | 37. |
| 3.8.3.1. Planteamiento general | 37. |
| 3.8.3.2. Máquinas de soporte no lineales y funciones kernel..... | 39. |
| 3.8.4. Multiclasificación con SVM..... | 41. |
| 4. CONSTRUCCIÓN DEL MODELO COMPUTACIONAL..... | 42. |
| 4.1. METODOLOGÍA..... | 42. |
| 4.2. RECOLECCIÓN DE MUESTRAS..... | 46. |
| 5. ADQUISICIÓN..... | 47. |
| 5.1. RELACIÓN PÍXEL-MICRA..... | 47. |
| 6. PRE-PROCESAMIENTO..... | 49. |
| 6.1. FILTRADO | 49. |
| 6.2. FILTRADO MORFOLÓGICO | 51. |
| 6.3. ELIMINACIÓN DEL RUIDO Y AUMENTO DEL CONTRASTE..... | 52. |
| 7. SEGMENTACIÓN | 55. |
| 7.1. UMBRALIZACIÓN..... | 55. |
| 7.2. CRECIMIENTO DE REGIONES | 56. |
| 7.3. SEGMENTACIÓN MANUAL | 57. |

| | |
|--|------------|
| 8. DESCRIPCIÓN..... | 59. |
| 8.1. MEDIDAS MORFOLÓGICAS..... | 59. |
| 8.1.1. Parámetros utilizados para describir el núcleo. | 59. |
| 8.1.2. Parámetros utilizados para describir el citoplasma..... | 61. |
| 8.2. MEDIDAS DE TEXTURA..... | 62. |
| 9. INTERPRETACIÓN Y RECONOCIMIENTO..... | 63. |
| 9.1. Selección de parámetros SVM..... | 64. |
| 9.1.1. Validación Cruzada..... | 65. |
| 9.1.2. Búsqueda en Malla..... | 65. |
| 10. RESULTADOS..... | 66. |
| 10.1. CLASIFICACIÓN DE CÉLULAS..... | 66. |
| 10.2. COMPARACIÓN RED NEURONAL SOFTWARE DIAGNOSIS VERSUS MAQUINAS DE SOPORTE VECTORIAL (SVM)..... | 68. |
| CONCLUSIONES Y RECOMENDACIONES..... | 71. |
| REFERENCIAS..... | 73. |
| ANEXOS..... | 74. |
| A. FUNCIONAMIENTO DE LA HERRAMIENTA SOFTWARE..... | 74. |

GLOSARIO

ADQUISICIÓN. Etapa del procesamiento digital de imágenes que permite obtener una imagen digital de un objeto, a partir de una escena.

ATIPIA CELULAR. Anormalidad en la forma o tamaño de las células.

BASOFÍLICO. Que presenta alto grado de coloración.

BINARIZACIÓN. Procedimiento para convertir una imagen que se encuentra en el formato de color verdadero, o en escala de grises, a un formato de dos colores (Blanco y Negro).

CÉLULAS ESCAMOSAS. Células encontradas en una muestra de citología cérvico uterina. De acuerdo con su etapa de maduración se clasifican en: basales, para basales, intermedias y superficiales.

CÉRVIX. Llamado también cuello uterino. Porción del útero que se une con la vagina.

CITOLOGÍA CÉRVICO UTERINA. Prueba de tamizaje realizada a las mujeres, que consiste en obtener una muestra adecuada de la unión exoendocervical y del endocérvix.

CIANOFÍLICO. Que tiene afinidad por los colorantes azules o verdes.

CITOLISIS. Disolución o destrucción celular.

CROMATINA. Sustancia protoplásmica, coloreable, del núcleo de la célula

DESECACIÓN CELULAR. Degeneración celular.

DESCRIPCIÓN. Etapa del procesamiento digital de imágenes que permite describir las características de un objeto.

DISCRETIZACIÓN. Procedimiento para asignar un valor discreto a una función continua.

DISPLASIA. Anomalía en el desarrollo de una célula.

DIGITALIZADOR. Dispositivo electrónico que permite convertir una señal análoga a un formato digital, por medio de la discretización.

ESONOFÍLICAS. Véase cianófilo.

HIPERCROMASIA. Abundancia de cromatina en una célula. En imágenes digitales se manifiesta con bajos niveles de color.

HIPERCROMATISMO. Véase hiper Cromasia.

HISTOGRAMA. Gráfico utilizado para la representación de la cada uno de los niveles de color.

LESIÓN INTRAEPITELIAL ESCAMOSA. Término acuñado por el sistema Bethesda para referirse a un cambio en el tejido del cérvix, a veces se usa como otra palabra para tumor.

NEOPLASIA INTRAEPITELIAL CERVICAL. Crecimiento anormal que se origina en el cérvix, causado por una sola célula alterada.

NIVEL DE COLOR. Valor asociado con la cantidad de luz detectada por un dispositivo de adquisición. Depende del número de bits empleado, para 8 bits se encuentra entre el intervalo [0,255].

ORANGOFÍLICO. Que presenta una coloración anaranjada.

PERINUCLEAR. Que rodea el núcleo celular

PICNOSIS. Condensación celular en la que el protoplasma se hace más denso, propia de una célula que se muere.

PLEOMORFISMO NUCLEAR. Variaciones morfológicas nucleares.

PRE-PROCESAMIENTO. Véase filtrado.

RADIO NÚCLEO/CITOPLASMA. Relación entre el tamaño del núcleo y el citoplasma.

RANGO DINÁMICO. Es el rango de diferencia tonal entre la parte más clara y la más oscura.

RECONOCIMIENTO: Etapa del procesamiento digital de imágenes que permite la clasificación de los objetos presentes en una escena.

SEGMENTACIÓN. Etapa del procesamiento digital de imágenes que permite la separación de los objetos presentes en una imagen.

SINCITIO. Célula simple o masa protoplasmática con muchos núcleos

SISTEMA BASADO EN EL CONOCIMIENTO. Sistemas informáticos que involucran conocimiento propio de especialistas.

TRANSFORMADA DE FOURIER. Función matemática que permite obtener las componentes de frecuencia de una señal.

UMBRALIZACIÓN. Técnica de segmentación que consiste en binarizar una imagen, tomando como referencia un nivel del histograma.

Índice de Figuras

| | | |
|-----|--|----|
| 1. | Células superficial | 7 |
| 2. | Células intermedias | 8 |
| 3. | Células para-basales | 9 |
| 4. | Espacio de color del modelo RGB | 15 |
| 5. | Escalas del modelo RGB | 15 |
| 6. | Espacio de color del modelo HSL | 16 |
| 7. | Escalas del modelo HSL | 17 |
| 8. | Etapas fundamentales del procesamiento de imágenes | 19 |
| 9. | División y fusión de regiones | 23 |
| 10. | Crecimiento de regiones | 24 |
| 11. | Representación de firmas | 26 |
| 12. | Hiperplanos que separan correctamente un conjunto de datos | 33 |
| 13. | Hiperplano lineal clasificador para el caso no separable | 38 |
| 14. | La SVM no lineal establece una relación del espacio de entrada con otro de representación de dimensión alta. | 39 |
| 15. | Diagrama de casos de uso tratamiento de imágenes | 43 |
| 16. | Diagrama de casos de uso SVM | 44 |
| 17. | Imagen digital con diagnóstico de malignidad | 47 |
| 18. | Cámara de Neubauer | 48 |
| 19. | Filtro promedio, gauss, circular y rotacional | 50 |
| 20. | Filtrado morfológico | 52 |
| 21. | Eliminación del ruido para imágenes con bajo contraste | 53 |
| 22. | Ajuste de intensidad | 54 |
| 23. | Eliminación del ruido utilizando ajuste del contraste | 54 |
| 24. | Crecimiento de regiones. Segmentación del núcleo | 56 |
| 25. | Crecimiento de regiones. Segmentación del citoplasma | 57 |
| 26. | Resultado de la segmentación manual del núcleo | 58 |
| 27. | Resultado de la segmentación manual de citoplasma | 58 |
| 28. | Núcleo detectado y diámetro calculado | 60 |
| 29. | Citoplasma detectado y diámetro calculado | 61 |
| 30. | Representación de las SVM | 64 |
| 31. | Datos de entrenamiento SVM | 69 |
| 32. | Datos de prueba SVM | 70 |
| 33. | Instalación de la toolbox SVM | 74 |
| 34. | Salvar toolbox SVM | 75 |
| 35. | Inicio de la herramienta software | 76 |
| 36. | Ventana principal de la herramienta software | 76 |
| 37. | Opción abrir imagen | 77 |
| 38. | Opción seleccionar | 78 |
| 39. | Ventana para el análisis del núcleo | 78 |
| 40. | Opciones de pre-procesamiento | 79 |
| 41. | Opciones de aumento del contraste 1 | 80 |
| 42. | Opciones de aumento del contraste 2 | 81 |
| 43. | Opciones de aumento del contraste 3 | 82 |
| 44. | Segmentación del núcleo | 83 |
| 45. | Identificar núcleo | 84 |

| | | |
|-----|---|----|
| 46. | Seleccionar núcleo | 84 |
| 47. | Núcleo segmentado | 85 |
| 48. | Datos de las características del núcleo | 85 |
| 49. | Eliminación del ruido para el citoplasma | 86 |
| 50. | Núcleo y citoplasma segmentados | 87 |
| 51. | Diagnóstico de la herramienta clasificadora | 87 |
| 52. | Diagnóstico de la herramienta clasificadora 1 | 88 |
| 53. | Ventana principal | 88 |
| 54. | Ventana configuración | 89 |
| 55. | Creando una nueva SVM | 89 |
| 56. | SVM creada y puesta en el sistema | 90 |
| 57. | Cargar SVM | 91 |
| 58. | Cambiar SVM | 92 |
| 59. | Cambio de la SVM | 93 |
| 60. | Ventana configuración 2 | 93 |
| 61. | Base de datos | 94 |
| 62. | Ventana configuración 3 | 94 |
| 63. | Carga de la SVM directamente | 95 |
| 64. | Archivo SVM cargado | 95 |

Índice de Tablas

| | |
|--|----|
| 1. Documento de la descripción de los casos de uso | 44 |
| 2. Resultados de las SVM. | 67 |
| 3. Datos software Diagnosis | 68 |
| 4. Datos de prueba y entrenamiento para las SVM y red Neuronal | 69 |

Índice de anexos

A. Funcionamiento de la Herramienta Software

74

RESUMEN

TÍTULO: HERRAMIENTA BASADA EN MÁQUINAS DE SOPORTE VECTORIAL PARA LA CLASIFICACIÓN DE CÉLULAS ESCAMOSAS DE CITOLOGÍAS CÉRVICO UTERINAS*

AUTORES:

LAURA YADIRA VALBUENA RODRÍGUEZ
MILTON ANTONIO CAVIEDES SANABRIA **

PALABRAS CLAVE: Tratamiento digital de Imágenes, Citología Cérvico Uterina, Máquinas de Soporte Vectorial, Células Escamosas.

El presente proyecto continúa el trabajo realizado por el grupo de Investigación en Ingeniería Biomédica (GIIB) en la línea de tratamiento de imágenes médicas, orientada a la construcción de aplicaciones futuras que permitan contribuir al diagnóstico de ciertas patologías que se apoyan en la lectura de Citologías Cérvico Uterinas y de tal forma reducir el error ocasionado por sobrevaloración o subvaloración de las características obtenidas de las células escamosas. La investigación ha sido enfocada a la caracterización y clasificación de células escamosas, según parámetros de anormalidad y normalidad utilizando la herramienta de máquinas de soporte vectorial para dicha clasificación.

Dicha herramienta de clasificación permite una mayor exactitud en el diagnóstico de las lecturas, ya que su diseño busca no solo encontrar una muy buena separación entre los diferentes elementos, sino la mejor separación que pueda existir entre los elementos a clasificar, dando de esta manera una mayor confianza al usuario final que el diagnóstico es correcto.

Se desarrollo una herramienta pensando en el usuario, que permita al experto no sólo obtener un software que le apoye al momento de tomar decisiones acerca de su trabajo, sino que le permita realizar de un manera fácil y rápida, la actualización de la base de datos, la inclusión o la eliminación de información de la misma, así como le permita a este elegir las características principales que se deban tener en cuenta durante el proceso de calificación, esta facilidad es de gran valor, ya que al trabajar con imágenes que tengan características similares, la posibilidad de un cambio en la calidad de estas afecta directamente el resultado que se espera tener.

Se recomienda que la continuidad de la investigación se dé en el análisis de campos que contengan un gran número de células y el uso de arquitectura en paralelo para la construcción de algoritmos más robustos

*Proyecto de Grado

**Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas
Director: Víctor Eduardo Martínez Abaunza
Codirectora: Olga Mercedes Álvarez Ojeda

ABSTRACT

TITLE: SOFTWARE BASED ON A SUPPORT VECTOR MACHINE FOR CLASIFICATION CELLS SQUAMOUS CELLS OF CERVICAL SMEAR CYTOLOGIES*

AUTHORS:

LAURA YADIRA VALBUENA RODRÍGUEZ
MILTON ANTONIO CAVIEDES SANABRIA**

KEYWORDS: Digital Image Processing, Cervical Smear Cytology, Squamous Cells, Support Vector Machine.

Description: The current project continues the work performed by the Biomedical Engineering Research Group (GIIB) in digital image processing, focusing on the construction of future applications which will allow to diagnose some pathologies that are supported on readings of cervical smear cytology, so that it is possible to decrease the errors made by the overestimation or the underestimation of distinguishing and classifying squamous cells according to the parameters of normality and abnormality, using the tool of support vector machine for such classification.

The known tool allows a greater accurateness on diagnosing the reading intended not just to reach a good distance among the different elements but the best distance that can be reached among elements to be classified, and on this way, providing the final user a major trust by certifying a correct diagnose.

Thinking on the users, a tool was developed, letting the expert not only obtain a software which backs up at the moment of making decisions about his or her job, but allowing him/her to perform a fast and easy way to update the database, including or getting rid of information itself as well as permitting this one to choose the main characteristics that are supposed to keep in mind during the process of qualification. This is of great worth because once it is working with images having similar characteristics, the possibility of a change in the quality of these make a direct result which it is expected to have.

The continuity in the research is recommended in analyzing fields which have a great amount of cells, and the use of parallel architecture for the construction of more robust algorithms.

* Research Work.

** Faculty of Engineering Physique-Mechanics. School of Systems Engineering and Informatics. Director: Víctor Eduardo Martínez Abaunza. Co-director: Olga Mercedes Álvarez Ojeda.

1. INTRODUCCIÓN

El cáncer de cuello de útero es la segunda causa de muerte por cáncer en mujeres con 450.000 nuevos diagnósticos cada año en el mundo. El 93% de las mujeres que lo padece están también infectadas por un virus de transmisión sexual, el papilomavirus. En Colombia se diagnostican aproximadamente 6.900 casos nuevos de cáncer de cérvix cada año y se presentan alrededor de nueve muertes al día por esta enfermedad. En Santander, en el año 2004 el tumor maligno de cuello uterino representó la tercera causa de mortalidad por cáncer en las mujeres con una tasa de 4,6 muertes por 100.000 mujeres [DECH09].

El 80% de las muertes por cáncer de cérvix se produce en países en desarrollo [DECH09].

Si todas las mujeres se hicieran la citología cérvico uterina, con regularidad, la mayoría de las condiciones precancerosas se detectarían y tratarían antes de que se volvieran cancerosas. De esa forma, se podría prevenir la mayoría de los cánceres invasores. Cualquier cáncer invasor que se presenta ahora, habría sido encontrado en una etapa temprana, curable.

Gracias al análisis de las citologías cérvico uterinas es posible detectar el cáncer de cerviz a tiempo, pero posee falencias, que son atribuidas en una tercera parte a errores en la lectura e interpretación de las muestras.

Al mejorar los sistemas de clasificación, los especialistas podrán tener mayor confianza en el resultado, disminuyendo la posibilidad de un falso positivo (anomalías, cuando en realidad no hay seguridad de ello) o falso negativo (normalidad, cuando tampoco hay seguridad de esta), resultados que indudablemente afectaran el factor anímico y emocional de las pacientes.

2. PRESENTACIÓN DEL PROYECTO

2.1 OBJETIVOS

2.1.1. Objetivo general

Desarrollar un software mediante la aplicación de un método basado en Máquinas de Soporte Vectorial para la clasificación de células escamosas en citologías cérvico uterinas, teniendo como base las características de normalidad existentes.

2.1.2. Objetivos específicos

- Actualizar las características existentes en las células escamosas de la zona cérvico uterina resultantes de la adquisición de imágenes.
- Complementar las características y expresarlas por medio de valores cuantificables y cualificables.
- Implementar la herramienta de máquinas de soporte de vectorial para la clasificación de los parámetros establecidos.
- Construir un software mediante el cual, el especialista obtenga información para la toma de decisiones acerca de la clasificación de las células escamosas en citologías cérvico uterinas.
- Comparar los resultados de la clasificación de las células realizada por una red neuronal versus la realizada por la herramienta de máquinas de soporte vectorial.

2.2. JUSTIFICACIÓN

Con este trabajo de grado, se busca implementar una nueva versión, que mediante la utilización de una herramienta basada en Maquinas de Soporte Vectorial (SVM) nos brinde mejores tiempos de respuesta y aciertos en la clasificación de células cervicales.

Es necesario, ampliar la caracterización de las imágenes de células cervicales para una mejor clasificación; adquirir imágenes de células grupales y no sólo individuales como hasta ahora se ha hecho.

Se utilizará la herramienta basada en Maquinas de Soporte vectorial (SVM), ya que esta técnica ha demostrado un gran desempeño en otros tipos de clasificación, además de su bajo costo de inversión, implementación y facilidad de manejo, por otro lado al aumentar la cantidad de características para clasificación, se espera mejorar su desempeño, logrando un alto grado de confiabilidad de los resultados obtenidos.

Por otro lado, la práctica ha demostrado que la adquisición de imágenes individuales es escasa, por lo tanto es necesario que el software tenga la facilidad de hacer el análisis con imágenes de células grupales.

3. MARCO TEÓRICO

3. 1. CITOLOGÍA CÉRVICO UTERINA

La citología cérvico uterina, se realiza para diagnosticar el cáncer cérvico uterino.

Esta prueba es un examen citológico en que se toman muestras de células epiteliales, en la zona de transición del cuello uterino, en busca de atipias celulares que orienten (no diagnostiquen) a la presencia de una posible neoplasia de cuello uterino. Las muestras para esta prueba se toman de tres partes: ENDOCERVIX, que es el orificio que comunica con el útero; CÉRVIX, que es la parte más externa del útero, y que comunica directamente con la Vagina; y el Tercio Superior de la Vagina, que es la parte que rodea el Cérvix o Cuello del Útero. Es importante recordar que durante el embarazo, la citología cérvico uterina no se toma de endocérvix, sino únicamente de cérvix y vagina.

Se practica generalmente a todas las mujeres desde que inician su actividad sexual (si no inician la actividad sexual a los 20 años, aproximadamente, igualmente se recomienda efectuar este estudio, porque la mujer podría contraer cáncer uterino) hasta los 65 años, aproximadamente.

Es un examen barato, fácil y simple, que cualquier médico puede realizar. Y también muy necesario para preservar la salud de la mujer. La citología obtenida debe ser analizada por médicos anátomo-patólogos experimentados [WIKI08].

3.1.1. Coloración de Papanicolaou

Se trata de una técnica citológica compuesta por 3 soluciones:

- Hematoxilina Biopur
- OG Biopur
- EA Biopur

3.1.2. Hematoxilina Biopur: Colorante nuclear.

- **OG Biopur:** Se trata de una solución alcohólica de ácido fosfotúngstico y Orange G. El ácido fosfotúngstico acidifica la solución y aumenta la unión del Orange G a zonas citoplasmáticas densas.
- **EA Biopur:** Se trata de una solución alcohólica. Tiñe selectivamente el citoplasma de células que tuvieron actividad metabólica poco antes de la recolección de la muestra. Tiñe de rosa a rojo las células escamosas superficiales, cilios, nucleolo y eritrocitos [BIOP07].

3.2. COMPONENTES CELULARES EN LA CITOLOGÍA CÉRVICO UTERINA

El diagnóstico citológico está basado en el análisis de las características morfológicas de las células exfoliadas del cérvix, que dependerán de la edad de la paciente, fase del ciclo menstrual e influencia de factores exógenos tales como inflamación, terapias, o estados patológicos. Las células epiteliales más numerosas son las células escamosas que se desprenden tanto del exocérvix como de la pared vaginal [MART04].

3.2.1. Células Cervicales

El cérvix uterino, de forma cilíndrica, posee una mucosa endocervical compleja, irregular. En las cercanías del orificio cervical externo, se observa la transición abrupta entre el epitelio cilíndrico simple que tapiza el canal cervical y el epitelio

plano estratificado que recubre la superficie de la porción vaginal del cérvix. Las células de las capas superficiales de éste epitelio están repletas de glucógeno [VIEF08]. Las células que se localizan en el epitelio plano estratificado se clasifican en superficiales, intermedias, basales y para basales.

Las coloraciones que se mencionan a continuación, son producto de la reacción celular producida por la prueba de Papanicolaou.

3.2.1.1. Células Superficiales

Las células superficiales son células epiteliales escamosas, maduras, por lo general poligonales. El citoplasma mancha en cualquiera de los dos grupos: cianófilico o eosinófilico, y el núcleo es picnótico. El criterio para describir las células superficiales, además de la madurez del citoplasma, es la evidencia de núcleo picnótico independiente de la reacción de manchado del citoplasma.

Las células superficiales de alto grado de madurez son planas y muestran más o menos líneas prominentes dentro del citoplasma. Esas líneas son aparentemente causadas por la formación de pre-queratina. La mayoría de las células superficiales tienen dobleces, algunas están hacinadas y muestran un citoplasma cianófilico translúcido. Éstas pueden mostrar una relativa prominencia en el contorno celular, la cual es producida por la fijación a las placas de vidrio, donde se toma la muestra. Entre las células superficiales maduras y las menos maduras están las células en etapa de transición, las cuales son identificadas por criterio e iniciativa del grado de madurez. Una inmensa variedad de combinaciones son vistas, como las planas, cianófilicas, células superficiales con pre-queratinización líneas o plegamiento, eosinófilicas con finos bordes, y así sucesivamente. Otra característica intracelular distintiva es la cierta madurez, y la presencia de células que se proliferan rápidamente en presencia de gránulos intracelulares.

La característica más importante de las células superficiales, está en la apariencia del núcleo celular. La mencionada en primer lugar tiene núcleo picnótico, mientras que la segunda tiene núcleo vesicular. Varios criterios morfológicos han sido propuestos para diferenciar dos tipos de núcleo. Pundel y Lichtfus sugirieron un criterio de medición del núcleo que dice lo siguiente: todo núcleo con diámetro menor que $6\mu m$ debe ser llamado picnótico. El procedimiento involucra el uso de varios tipos de aparatos de medición celular y es dispendioso. El microscopio de contraste de fases puede ser empleado en el teñido de especímenes. En este microscopio una desviación óptica del camino de la luz causa que los cuerpos más densos resalten con un rojo brillante, mientras que el núcleo vesicular aparece más oscuro y opaco [DISO05].

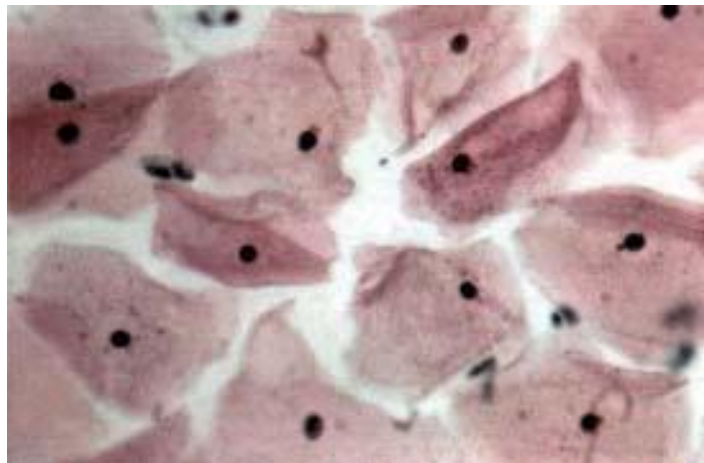


Figura 1. Células Superficiales

3.2.1.2. Células Intermedias

Las células intermedias son células escamosas epiteliales maduras que pueden ser tanto eosinófilicas o cianófilicas pero que contienen núcleo vesicular no picnótico. Las células intermedias tienen relativamente, un gran desarrollo celular epitelial comúnmente tan grande como las células

superficiales, pero que no pueden ser confundidas con una célula superficial porque su núcleo no es picnótico. Tampoco pueden confundirse con células para-basales porque esta a menudo es pequeña, redonda u ovalada. Las células intermedias tienden a presentar plegamiento celular, y con frecuencia hacinamiento entre grupos de células pequeñas o grandes. Sin embargo las formas más maduras de las células intermedias pueden también ser planas y solitarias.

Al igual que las células superficiales, existe una gran variedad de combinaciones para hacer posible la distinción entre los grados de madurez en que se encuentran células intermedias como por ejemplo células plegadas esonofílicas que se encuentran solitarias o células planas cinéfilas que tienden a formar grupos. Esto llega a ser obvio cuando se requiere más de un criterio para describir el patrón correcto cuando se desea describir cierta célula epitelial [DISO05].



Figura 2. Células Intermedias

3.2.1.3. Células Para-Basales

A menudo las características de las Este grupo abarca las células pequeñas, ovaladas o redondas, inmaduras, en células epiteliales escamosas. Estas células generalmente contienen gran estructura, rara vez núcleo picnótico degenerativo. Las células para-basales no están definidas; pero están definidas

para el grupo que abarca todas las células normales, fisiológicamente ocurre en las células escamosas epiteliales, que no son superficiales ni intermedias. La reacción al manchado es comúnmente cianofílica o poco clara, rara vez eosinofílica. El término para-basal fue escogido porque este representa un tipo de célula que no se identifica en la actual capa epitelial de la cual son derivadas. Parece altamente improbable que alguna citopatología pueda ser capaz de identificar una verdadera célula basal únicamente por la observación de su cito morfología. Es altamente concebible que al exfoliar células para-basales, que puedan ser derivadas de la capa superficial de un muy desarrollado epitelio que es inmaduro, no estén claramente diferenciadas dentro de las varias capas de células del epitelio. Las células para-basales no muestran citólisis bacterial, como se observa en las células intermedias. Algunas muestran degeneración auto lítica en el sentido de una separación de proteínas necrótica [DISO05].

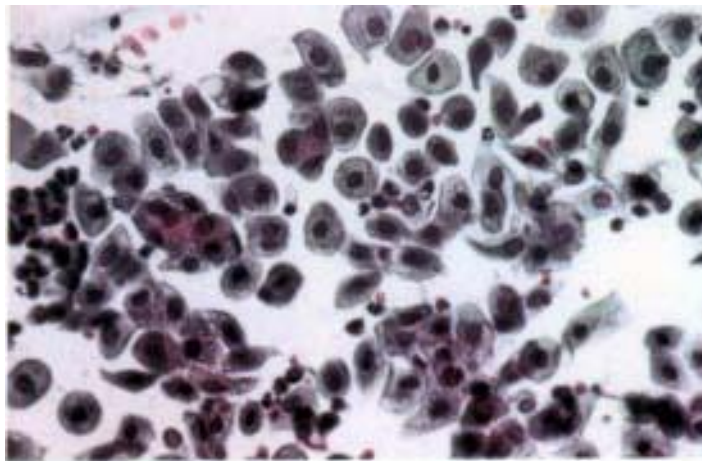


Figura 3. Células Para-Basales

3.3. ESTRUCTURAS CELULARES QUE SON PARÁMETROS DE EVALUACIÓN

Para la evaluación tanto de actividad biológica y funcional de las células se deben tener en cuenta los siguientes parámetros:

- Núcleo.
- Nucleolo.
- Cromatina.
- Borde cromatínico y membrana nuclear.
- Red de cromatina.
- Paracromatina.
- Mitosis.
- Multinucleación.
- Radio núcleo/citoplasma.
- Coloración del citoplasma

3.4. PROCESOS PATOLÓGICOS QUE AFECTAN LAS CÉLULAS

3.4.1. Efectos morfológicos sobre células que resultan en una actividad normal. Euplasia.

En este estado las estructuras clave vistas al microscopio de luz son: núcleo, nucleolo, cromocentro redondeados, y un patrón de cromatina y paracromatina uniforme con un grosor regular del anillo cromatínico asociado internamente a la membrana nuclear [MART04].

3.4.2. Efectos morfológicos sobre células de procesos asociados con cáncer.

Algunos cambios morfológicos claves son: membrana nuclear de forma irregular. Hay ángulos hacia adentro y hacia afuera del núcleo. La cromatina se condensa en grumos gruesos que se distribuyen irregularmente, principalmente cerca del anillo cromatínico. La paracromatina es fácilmente apreciable, ocasionada por la mayor cantidad de zonas claras. El nucleolo es grande, irregular y de forma angulada. Con mitosis anormales. Multinucleación con núcleos pleomórficos. Incremento del radio núcleo/citoplasma [MART04].

3.4.3. Efectos morfológicos sobre células de procesos retrogresivos resultando en una disminución de la actividad. Retroplasia.

Hay edema celular dado por degeneración hidrópica o crenación, cambio en la coloración de tinción citoplasmática, núcleo arrugado, colapsado, disminución de la basofilia nuclear, cromatina agregada en grumos gruesos con apariencia de gelatina [MART04].

3.4.4. Efectos morfológicos sobre células de procesos progresivos resultando en actividad aumentada. Proplasia.

La membrana nuclear muestra ondulaciones, hipercromasia nuclear, mayor aclaramiento de la paracromatina distribuida regularmente, nucleolos inmensos y en ocasiones múltiples, con esquinas redondeadas, mitosis normales, multinucleación frecuente con regularidad nuclear, y una relación núcleo/citoplasma aumentada [MART04].

3.4.5. Efectos morfológicos sobre células de procesos cambiantes y mixtos.

Pueden mezclarse los eventos anteriormente descritos [MART04].

3.5. DIAGNÓSTICO EN CITOLOGÍA

Para analizarla se utiliza la clasificación del Sistema Bethesda, esta tiene como complemento informar sobre lo adecuado de la muestra, y un diagnóstico descriptivo, con reconocimiento de patrones morfológicos permitiendo discernir entre cambios celulares benignos por infecciones o reparativos y anomalías de las células epiteliales. En esta última categoría se debe clarificar si los cambios observados pertenecen a condiciones premalignas o malignas y en caso dudoso se establece la categoría de *SIGNIFICANCIA INDETERMINADA* [MART04].

En el diagnóstico descriptivo hay áreas que generan gran controversia particularmente las denominadas células escamosas o glandulares atípicas de significancia indeterminada. (ASCUS – AGUS). Se usa cuando los cambios encontrados exceden a los ocasionados por proceso reactivo benigno, pero son insuficientes para un diagnóstico definitivo de lesión. La *LESIÓN INTRAEPITELIAL ESCAMOSA (LISL)* se divide en lesión de bajo y alto grado [MART04]. Hay características morfológicas relacionadas con cada una de estas:

- En el diagnóstico por virus del papiloma humano (VPH) hay cambios nucleares que incluyen agrandamiento nuclear con hipercromasia o picnosis, con cromatina y contorno nuclear arrugado, en el citoplasma hay una cavidad perinuclear bien definida con reforzamiento del citoplasma adyacente [MART04].
- Las células de la displasia leve pertenecen a la categoría de bajo grado, muestran un citoplasma traslúcido, con bordes angulados bien definidos, recordando células superficiales o intermedias pero con menor cantidad de citoplasma y un núcleo más grande que ocupa menos de un tercio del área total de la célula. La cromatina nuclear es finamente granular, irregularmente distribuida con ligera hipercromasia [MART04].
- En las de alto grado se incluyen la displasia moderada, severa y el carcinoma in situ. En la displasia moderada las células tienen un tamaño variable, la mayoría de ellas son ovales, pero pueden encontrarse células fusiformes bizarras. Predominan los citoplasmas eosinofílicos con núcleos agrandados de forma irregular. Hay mayor aumento en el radio núcleo/citoplasma. El nucleolo usualmente está ausente. En la displasia severa y carcinoma in situ el tamaño de las células se compara con las células para basales, mostrando gran

núcleo que ocupa dos tercios del área total de la célula con hipercromasia, cromatina en grumos gruesos distribuidos irregularmente y membrana nuclear con irregularidades [MART04].

- Cuando la lesión tumoral sobrepasa la membrana basal en donde se apoya el epitelio escamoso, es considerada como invasora, siendo muy agresiva. Existen parámetros definidos para el diagnóstico de carcinoma invasor como son: menor tamaño de las células escamosas tumorales, comparadas con las células normales con formas redondas, ovals y alargadas conteniendo fibrillas intracitoplasmáticas. Las células se disponen en sincitios con bordes no definidos, con polaridad alterada. El núcleo tiene pleomorfismo, con un alto radio núcleo/citoplasma [MART04].

3.5.1. Sistema Bethesda 2001

El sistema Bethesda comprende cambios basados en la experiencia clínica y en avances del conocimiento acerca de la biología del carcinoma cervical. El diagnóstico final y el esquema de tratamiento de un paciente contemplan no sólo el resultado del estudio citológico sino también la anamnesis, los hallazgos clínicos y otros estudios de laboratorio, así como las interpretaciones de la histología.

El sistema fue elaborado principalmente para clasificar muestras citológicas cervicales, lo cual se ve reflejado tanto en la terminología como en los criterios morfológicos [DISO05].

3.6. TRATAMIENTO DIGITAL DE IMÁGENES

3.6.1. Fundamentos

El interés por los métodos de tratamiento digital de imágenes deriva de dos áreas principales de aplicación: la mejora de la información pictórica para la interpretación humana y el procesamiento de los datos de la escena para la percepción autónoma por una máquina, para el segundo caso, el interés se centra en los procedimientos para extraer la información de la imagen de forma conveniente para el procesamiento por computadora. A menudo esta información tiene poco en común con los rasgos visuales que los seres humanos emplean para interpretar el contenido de una imagen.

El advenimiento combinado de las computadoras digitales de gran potencia y la carrera espacial fue lo que puso de manifiesto el potencial de los conceptos del tratamiento digital de imágenes (TDI) de ahí en adelante el TDI representó un gran avance y ayuda, no solo en el área espacial sino también en biología, medicina nuclear, investigaciones judiciales, defensa y aplicaciones industriales [GOWO96].

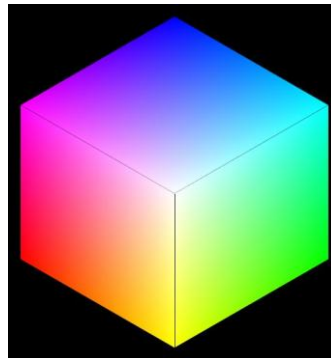
3.6.2. Modelos de Color

El objetivo de un modelo de color es facilitar la especificación de los colores de una forma normalizada y aceptada genéricamente. La mayoría de modelos de color empleados están orientados hacia el hardware, otros hacia aplicaciones que pretenden manipularlo. Por esto, se hará una introducción a los espacios de color más comunes: el modelo RGB (Red, Green, Blue), para el primer caso; y el modelo HSL (Hue, Saturation, Luminance), para el segundo caso [GOWO96] [MART07].

3.6.3. Modelo de color RGB

El color aparece con sus componentes espectrales primarias de rojo, verde y azul. Basado en un sistema de coordenadas cartesianas. En la figura 4 se observa que el subespacio de color de interés es un cubo. En este modelo los colores son definidos por todos los vectores que se extienden desde el origen hasta cualquier punto del cubo [GOWO96] [MART07].

Figura 4. Espacio de color del modelo RGB



Las imágenes de este modelo consisten en tres planos independientes, uno por cada color primario, los cuales se superponen para formar una imagen en color. La figura muestra las variaciones que se presentan en los colores primarios del modelo RGB. Cuando la proporción de colores es igual en sus tres componentes se obtiene la escala de grises [GOWO96] [MART07].

Figura 5. Escalas del Modelo RGB



(a) Escala de Rojo



(b) Escala de Verde



(c) Escala de Azul

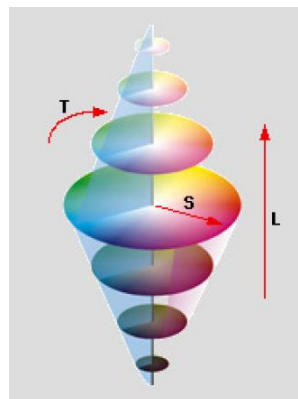


(d) Escala de Grises

3.6.4. Modelo de color HSL.

En este modelo, el color se representa como dos conos unidos por su base (figura 6). La componente H representa el matiz del color (rojo, amarillo, verde, cian, azul, magenta) y está definida como un ángulo; La componente S o saturación es la cantidad del matiz que ha sido diluido en blanco. Mientras que la componente L es la escala de grises o Luminancia. En este modelo las componentes del matiz y saturación están íntimamente relacionadas con la forma en que los seres humanos perciben el color [GOWO96] [MART07].

Figura 6. Espacio de Color del Modelo HSL

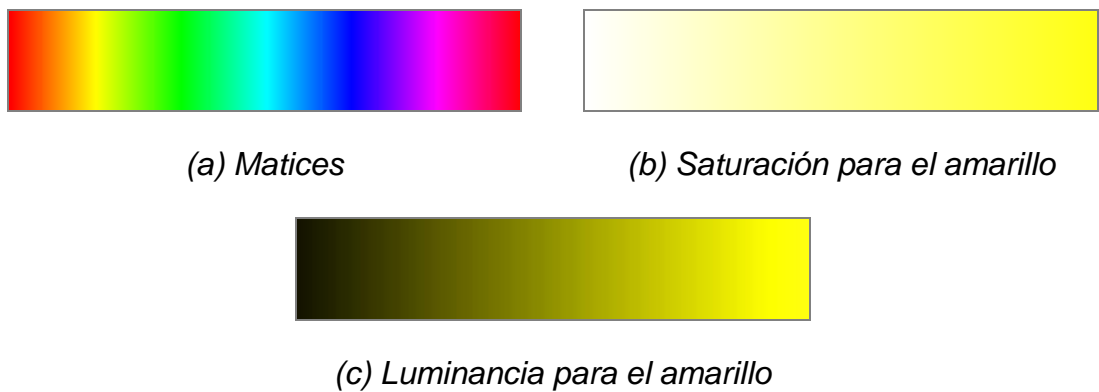


Fuente: HANBURY, Allan. Technical Report N-04/01/MM. Paris: Centre de Morphologie Mathématique. Ecole de Mines de Paris, 2001. p. 7.

En la siguiente figura se observa una escala de matices en la que se distinguen tonos naranjas, amarillos, verdes, azules, etc. En la escala de saturaciones para un estímulo amarillo se observa como varía el tono en

función de la saturación. En la escala de luminosidades, el matiz emite más o menos luz de acuerdo con la modificación que se haga en los valores [GOWO96] [MART07].

Figura 7. Escalas del Modelo HSL



3.6.5. Conversión del Modelo RGB a HSL.

La representación del espacio RGB en términos de coordenadas polares 3D es usada frecuentemente en el análisis de imágenes. Aunque presenta algunas fallas, ofrece una representación más intuitiva, y revela características de la imagen que no son visibles con ese modelo. La transformación entre estos dos sistemas es una simple operación de coordenadas [GOWO96] [MART07].

Las siguientes funciones proporcionan los valores HSL en el intervalo [0,1]. Se debe cumplir también que las componentes RGB se encuentren normalizadas en el mismo intervalo [GOWO96] [MART07]. La función para el matiz es:

$$H = \cos^{-1} \left\{ \frac{\frac{1}{2} \left[\max(R-G, R-B) - \min(R-G, R-B) \right]}{\left[\max(R-G, R-B) + \min(R-G, R-B) \right]} \right\} \quad (1)$$

Para la saturación se tiene:

$$S = 1 - \frac{3}{R+G+B} \min(R, G, B) \quad (2)$$

Finalmente, para la luminancia

$$L = \frac{1}{3} (R+G+B) \quad (3)$$

Sin embargo, en esta transformación se encuentra que:

- La saturación no necesariamente tiene un bajo valor para píxeles acromáticos.
- La información cromática y de brillo no son independientes. La función de luminancia es usada para normalizar la saturación, así que los valores de saturación dependen en gran medida de la función brillo.

Las anteriores desventajas deben solucionarse entonces, dentro de una nueva alternativa se tiene que las cantidades máx.-mín. asociadas a la saturación permiten una mejora del espacio HLS. Sin embargo, esta semi-norma no puede utilizarse para el brillo o luminancia, así que para ésta se utiliza la ecuación recomendada por el estándar de televisión de alta definición. Obteniéndose un sistema HLS mejorado o Improve Hue Luminance Saturation System (IHLS). El cual está fundamentado en las siguientes ecuaciones [GOWO96] [MART07]:

$$H' = \cos^{-1} \left\{ \frac{\frac{1}{2} \max(R-G, G-R)}{\max(R-G, G-R) + \max(R-B, B-R) + \max(G-B, B-G)} \right\} \quad (4)$$

$$H_{IHSL} = \begin{cases} 360^\circ - H' & \text{si } B > G \\ H' & \text{en otro caso} \end{cases} \quad (5)$$

$$S_{IHLS} = \text{máx}(R, G, B) - \text{mín}(R, G, B) \quad (6)$$

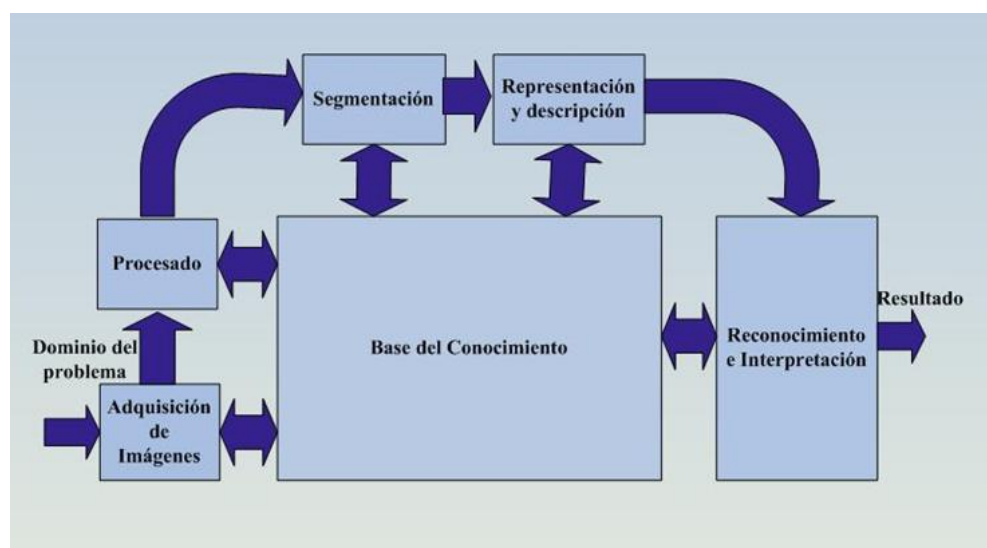
$$L_{IHLS} = 0.213R + 0.715G + 0.072B$$

Estas nuevas ecuaciones permiten separar la información del color de la información lumínica [GOWO96] [MART07].

3.7. ETAPAS DEL TRATAMIENTO DIGITAL DE IMÁGENES

Básicamente el tratamiento de imágenes se ocupa de las siguientes etapas: Adquisición, Pre-procesamiento, Segmentación, Descripción y Reconocimiento.

Figura 8. Etapas fundamentales del procesamiento de imágenes.



Fuente: González Rafael C. y Woods Richard E. Tratamiento Digital de Imágenes. Wilmington: Addison-Wesley Iberoamericana, 1996.

3.7.1. Adquisición

La primera etapa en el proceso es la adquisición digital de la imagen. Para ello es necesario la utilización de un dispositivo (cámara) sensible a una determinada banda del espectro de energía electromagnética (espectro visible) y que produzca una señal eléctrica de salida proporcional al nivel de energía detectado. Si la salida del dispositivo no está en formato digital se debe usar un conversor analógico digital (tarjeta digitalizadora). Luego de adquirir los datos digitalizados, éstos son almacenados en unidades de memoria. Una vez obtenida la imagen digital, el siguiente paso consiste en el pre-procesamiento de dicha imagen [GOWO96] [FEYE01].

3.7.2. Pre-procesamiento

En el proceso de adquisición la imagen que se genera no es totalmente confiable, dado que se pueden presentar obstáculos; fallos en la iluminación del microscopio, partículas de polvo en los objetivos, insuficiente coloración de las muestras, etc. El filtrado de una imagen permite corregir algunos de estos problemas y ha sido agrupado conforme a la técnica utilizada. Cada filtro es implementado de acuerdo a la imagen que se desea obtener [GOWO96] [MART04].

- **Mejoramiento por procesamiento de punto.** Se basa en la mejora de las intensidades de los niveles de grises considerando los píxeles individualmente. Entre los más utilizados están: Negativos de imágenes, Aumento del Contraste, Compresión del Rango Dinámico y Procesamiento del Histograma (Ecuilización y Especificación) [GOWO96] [MART04].

- **Filtrado en el Dominio Espacial.** El filtrado espacial de una imagen es implementado mediante pequeñas matrices cuadradas, conocidas como máscaras, núcleos o ventanas, las cuales, aplicando el teorema de la convolución con la imagen a filtrar, produce como salida otra imagen donde ha sido modificado el contenido en escala de grises. A su vez han sido clasificados como Paso-Bajo y Paso-Alto, los primeros permiten homogenizar las superficies de la imagen mediante el difuminado, los segundos son útiles en la extracción de bordes. Entre los filtros espaciales más comunes encontramos: El Filtro Mediana, El Filtro Promedio y los Filtros Diferenciales (Operadores de Roberts, Prewitt y Sobel) [GOWO96] [MART04].
- **Filtrado en el Dominio de la Frecuencia.** Para implementar esta serie de filtros se calcula la Transformada de Fourier de una imagen, con el fin de obtener las componentes frecuenciales de la imagen para multiplicarlas por una función de transferencia de un filtro, seguidamente aplicamos la transformada Inversa de Fourier para producir la imagen filtrada. Estos se encuentran clasificados como Paso-Bajo y- Paso-Alto, los primeros atenúan las componentes de alta frecuencia y los segundos las bajas [GOWO96] [MART04].

La función de transferencia más utilizada es conocida como Filtro de Butterworth tanto para frecuencias altas como para frecuencias bajas [GOWO96] [MART04].

3.7.3. Segmentación

La segmentación subdivide una imagen en sus partes constituyentes u objetos (Núcleo, Citoplasma, etc.). El nivel al que se lleva a cabo esta subdivisión

depende del problema a resolver. Esto es, la segmentación deberá detenerse cuando los objetos de interés de una aplicación hayan sido aislados [GOWO96].

Los algoritmos de segmentación pueden basarse en la discontinuidad o en la similitud de los píxeles; en el primer caso, las principales áreas de interés son la detección de puntos aislados y la detección de bordes y líneas, en el segundo es usual utilizar algoritmos como la umbralización o aquéllos orientados a regiones (crecimiento o división y fusión), dado que agrupan las píxeles de acuerdo con propiedades semejantes [GOWO96].

Umbralización: Es la técnica más común para la segmentación de imágenes, dado el bajo consumo de computación que requiere. Se basa en la suposición que los objetos producen regiones con niveles de luminosidad semejantes y al ser representados en un histograma mostrarán modas dominantes separadas por un nivel denominado umbral. Se tiene entonces que es una operación de la función T de la forma [MART07]:

$$U = T [f(x,y), p(x,y)] \quad (7)$$

Donde $f(x,y)$ es el nivel del punto (x,y) , y $p(x,y)$ representa alguna propiedad local de este punto. Una imagen umbralizada $g(x,y)$ se define como [MART07]:

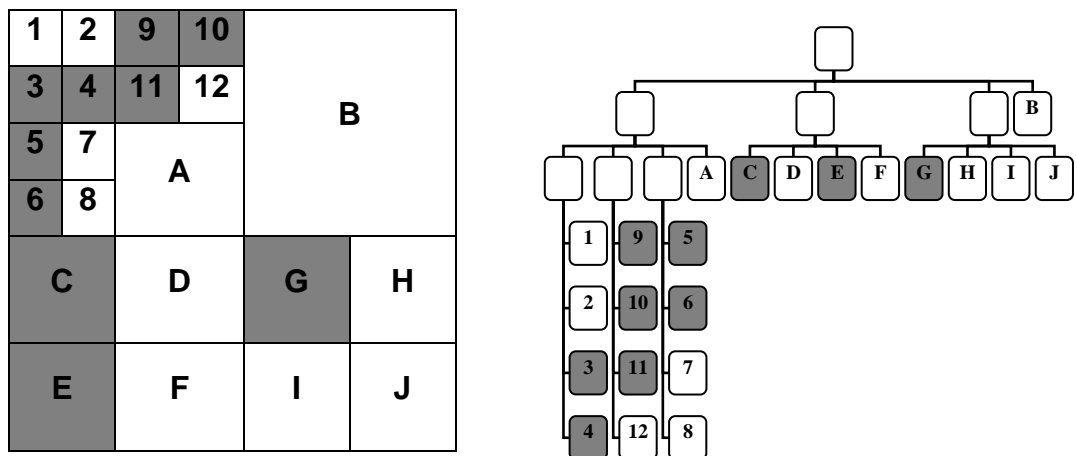
$$g(x,y) = \begin{cases} 1 & \text{si } f(x,y) > U \\ 0 & \text{si } f(x,y) \leq U \end{cases} \quad (8)$$

División y fusión de regiones: Es un método que consiste en dividir la imagen en regiones uniformes, como lo muestra la figura 9. El algoritmo comienza con la hipótesis que la imagen completa es una sola región, entonces analiza la homogeneidad de la misma (mediante algún cierto criterio soportado en las propiedades de la región). Si existe homogeneidad, la región se

encuentra ya segmentada; si no es así, entonces la región es dividida en cuatro regiones adicionales. Este proceso se repite para cada una de las regiones generadas hasta que la división no puede llevarse a cabo [MART07].

Luego de dividir la imagen, se comprueba si cada región generada se puede unir con una región adyacente (se evalúa nuevamente si satisface el criterio de homogeneidad establecido). El proceso termina cuando no se pueden fusionar más regiones [MART07].

Figura 9. División y fusión de regiones



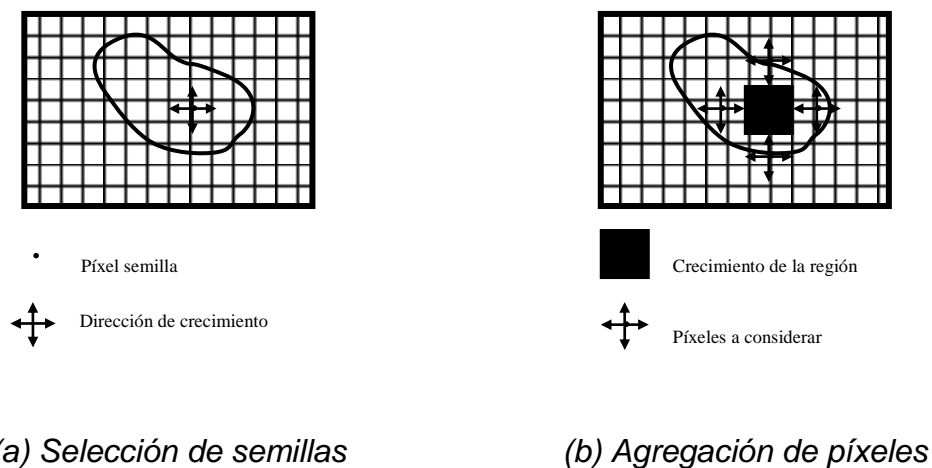
Fuente: González Rafael C. y Woods Richard E. Tratamiento Digital de Imágenes. Wilmington: Addison-Wesley Iberoamericana, 1996.

Crecimiento de regiones: Es un método de segmentación que utiliza un principio opuesto a la división y fusión. En la figura 16 se observa el resultado de este método, donde las regiones crecen mediante agregación de píxeles similares en valor, respecto a una propiedad de homogeneidad que se utilice para realizar la segmentación [MART07].

Este algoritmo inicia con la selección de un conjunto de puntos semilla en la imagen. Estos puntos semillas servirán como puntos de comienzo del proceso de crecimiento de las regiones, con lo cual, el número máximo de regiones finales es igual al número de semillas seleccionadas (aunque llegará a ser menor si dos regiones se unen para formar una sola) [MART07].

Para realizar la agregación de píxeles similares será necesario definir el concepto de similaridad, que puede no ser el mismo para todas las aplicaciones. Posibles criterios, ya utilizados en algoritmos desarrollados, son la diferencia entre el valor del píxel a agregar y el valor de la semilla, o que el valor medio de la región ya formada sea menor que un cierto umbral predeterminado [MART07].

Figura 10. Crecimiento de regiones



Fuente: GONZÁLEZ PENEDO, Manuel. Op. cit.

El comienzo con un píxel semilla permite que esa región crezca completamente; sin embargo, antes de tratar otras semillas se deben tener en cuenta varios aspectos [MART07]:

- En el crecimiento de regiones dominantes las ambigüedades respecto a los bordes de regiones adyacentes pueden no ser resueltas correctamente.
- La elección de diferentes puntos semilla puede dar lugar a diferentes segmentaciones.
- Pueden surgir problemas si un píxel semilla elegido aleatoriamente pertenece a un borde [MART07].

Para tratar de resolver estos problemas, se han desarrollado técnicas de crecimiento de regiones de forma simultánea:

- No se permite que una única región domine completamente el proceso.
- Un cierto número de regiones crecen al mismo tiempo. (regiones similares, presentan un mismo comportamiento en su crecimiento) [MART07].

3.7.4. Descripción

Una vez que han sido separados los objetos de la imagen, es necesario representarlos y describirlos mediante el uso de ecuaciones y modelos matemáticos, con el propósito de encontrar las propiedades que los identifiquen. Existen dos posibilidades para describir las regiones, en términos de sus características externas relacionadas con la forma (contornos y bordes) o por medio de los píxeles que comprenden el interior del objeto (texturas) [MART07].

- **Descriptores de forma** Es frecuente que algunos objetos puedan ser diferenciados de otros por su forma, la cual puede ser usada independientemente o en combinación con medidas relacionadas con el

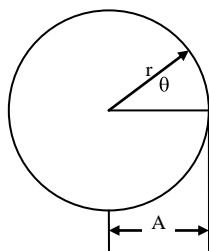
tamaño. Una medida muy común para la descripción de la forma es la circularidad, definida de la siguiente manera [MART07]:

$$C = \frac{P^2}{A} \quad (9)$$

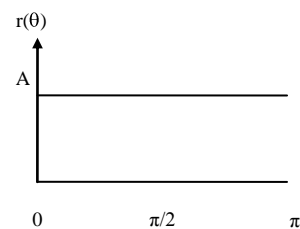
Es decir, es la relación entre el cuadrado del perímetro, distancia entre los píxeles del contorno, y el área, número de píxeles dentro del objeto. Esta medida toma el valor de 4π para una figura de forma circular [MART07].

Otro descriptor utilizado ampliamente en la representación de formas es la firma, y consiste en una representación unidimensional del contorno. El método más usual para determinar la firma es la distancia desde el centroide al contorno como una función del ángulo. La figura muestra una forma geométrica con su respectiva firma [MART07].

Figura 11. Representación de firmas



(a) Forma circular



(b) Firma de un círculo

De acuerdo con lo anterior, la firma puede ser expresada entonces como:

$$r = \sqrt{(x_c - b_{ix})^2 + (y_c - b_{iy})^2} \quad (10)$$

Donde c es la posición centroide y b es la posición de un píxel del borde.

- **Descriptores de textura** Esta propiedad de las regiones hace referencia a un atributo que representa el orden o las características espaciales de un objeto, y se busca obtener un valor que cuantifique las variaciones en los niveles de la imagen. Se espera que este descriptor sea independiente de la posición, orientación, tamaño, forma, y el nivel promedio del objeto [MART07].

Un descriptor utilizado para la representación de texturas es la entropía, definida por la ecuación [MART07].

$$\text{entropía} = -\sum_{i=0}^{255} p_i \log_2 p_i \quad (11)$$

Donde, p es el valor del histograma para el nivel $i \in [0,255]$.

También se cuenta con la matriz de co-ocurrencia, que se ha convertido en un modelo bastante útil en la representación de texturas construyéndose a partir de las veces que dos niveles ocurren en una distancia determinada, así: supóngase que se establece una dirección (horizontal, vertical, etc.) y una distancia (un píxel, dos píxeles, etc.) en una imagen; entonces el elemento i,j de la matriz de co-ocurrencia P es el número de veces, dividido por M , que el nivel i y j ocurren en dos píxeles separados por la distancia y la dirección seleccionada, donde M es el número total de pares de píxeles posibles. La matriz P es de tamaño $N \times N$, para una imagen de N niveles [MART07].

Debido a que el número total de pares de píxeles M es menor al número de píxeles del objeto, y disminuye si se incrementa la distancia, es usual reducir el número de niveles de 256 a 8, para el cálculo de la matriz de co-ocurrencia. Una vez determinada la matriz, se calculan los respectivos descriptores a partir de las siguientes ecuaciones [MART07]:

$$\text{contraste} = \sum_{i,j=0}^{N-1} |i-j|^2 p(i,j) \quad (12)$$

Donde $p(i,j)$ es la probabilidad de co-ocurrencia de los valores i,j para un distancia dada y N es el tamaño de la matriz. El contraste es cero para una textura homogénea y un valor alto en caso contrario [MART07].

$$\text{correlación} = \sum_{i,j=0}^{N-1} \frac{(i-\mu_i)(j-\mu_j) p(i,j)}{\sigma_i \sigma_j} \quad (13)$$

El rango de la correlación se encuentra en el intervalo $[-1,1]$. Indica como se encuentra un píxel respecto a toda la región y existen dos consideraciones importantes: un objeto tiene mas alta correlación dentro de él que entre objetos adyacentes y los píxeles cercanos están mas correlacionados entre sí que los píxeles mas distantes [MART07].

$$\text{energía} = \sum_{i,j=0}^{N-1} p(i,j) \quad (14)$$

La energía es una medida de la homogeneidad local, se encuentra en el intervalo $[0,1]$, y llega a su máximo valor para las regiones homogéneas [MART07].

$$\text{homogeneidad} = \sum_{i,j=0}^{N-1} \frac{p(i,j)}{1+|i-j|} \quad (15)$$

Esta medida es alta cuando la matriz de co-ocurrencia se concentra a lo largo de la diagonal. Esto ocurre cuando la imagen es localmente homogénea de acuerdo al tamaño de la ventana y posee un comportamiento similar a la energía [MART07].

- **Descriptores estadísticos** Debido a su generalidad, estos últimos pueden ser utilizados tanto para la representación de contornos, como para las regiones. Se recurre a la estadística descriptiva gracias a que sencillos modelos, como el promedio, ecuación 16; la desviación estándar, ecuación 17; la varianza, cuadrado de la desviación estándar; o la mediana, logran representar algunas características significativas de los objetos. Las ecuaciones aplicables son las siguientes [MART07]:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (16)$$

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (17)$$

3.7.5. Interpretación y Reconocimiento

En esta etapa se asigna una etiqueta a los objetos de acuerdo con sus descriptores permitiendo reconocer e identificar patrones de los objetos para ser comparados con otros patrones y poder así reconocer e identificar el objeto de la imagen que se ha analizado. Mediante clasificadores e la mínima distancia, de correspondencia por relación y los clasificadores bayesianos, el proceso que se ha aplicado a la imagen permite establecer criterios de decisión [GOWO96] [FEYE01].

Una herramienta que está implementándose en el reconocimiento de objetos son las Máquinas de Soporte Vectorial o SVM (en sus siglas en inglés Support Vector Machine).

En épocas recientes, se ha desarrollado una rama de la Inteligencia Artificial denominada aprendizaje supervisado [KUAH94], el cual se basa en un conjunto de datos (elementos de aprendizaje), de antemano conocidos, para la construcción de un modelo que describa un sistema dado y que permita la posibilidad de predicción ante datos desconocidos. En este caso, el elemento de aprendizaje está dando el valor correcto de la función correspondiente a determinadas entradas, y dicho elemento modifica la representación disponible de la función tratando de que corresponda a la información que le proporciona la retroalimentación. De manera formal, se dice que un ejemplo es un par $(x, f(x))$, en donde x es la entrada y $f(x)$ es la salida de la función que se aplicó a x . Entonces el objetivo es, dado un grupo de ejemplificaciones de f , producir una función h que se aproxime a f . A la función h , se le conoce como hipótesis.

Las máquinas de vectores de soporte o SVM son una consecuencia práctica de la teoría de aprendizaje y su estudio es útil básicamente por dos razones:

1. Desde el punto de vista teórico satisface totalmente la teoría de vectores de soporte para aprendizaje, basado en ideas maravillosamente simples y que proveen una clara intuición de que se está aprendiendo de los ejemplos.
2. Ofrece grandes resultados en aplicaciones prácticas. En este sentido se considera que el algoritmo de soporte vectorial (SV) es la intercepción entre la teoría y la práctica.

3.8. MÁQUINAS DE SOPORTE VECTORIAL (SVM)

Son una clase específica de algoritmos caracterizados por:

- Uso de funciones kernel
- Ausencia de mínimos locales
- Capacidad de control obtenida al manejar el margen.

Desarrollado por Vladimir Vapnik y sus colaboradores, su primera presentación fue en 1992. Aunque sus propiedades ya existían y habían sido usadas en las máquinas de aprendizaje desde 1960 [MORA05]:

- Los hiperplanos de margen grande en el espacio de entrada fueron tratados por Duda y Venado, Cubra, Vapnik, y otros.
- El uso de kernels se propuso por Aronszajn, Wahba, Poggio, y otros, pero fue Aizermann en 1964 quien introdujo la interpretación geométrica de los kernels como los productos internos en un espacio característico.
- El uso de variables de relajación para superar el problema de ruido y la no separabilidad, se introdujo en los años sesenta por Smith y perfeccionado por Bennett y Mangasarian [MORA05].

Sin embargo, no fue hasta 1992 que todas estas características se reunieron para formar el clasificador del margen fuerte, la Máquina Soporte Vectorial básica, y hasta 1995 la versión del margen débil se introdujo por Cortés y Vapnik: *“Es sorprendente como natural y elegantemente todas las piezas se ajustan entre si y se complementan unas con otras”*.

Las publicaciones de Shawe-Taylor y Bartlett dieron el primer límite estadístico riguroso en la generalización de las SVMs de margen fuerte. Luego en trabajos Shawe-Taylor y Cristianini dan límites similares para los algoritmos de margen débil.

Después de la introducción de las SVMs, un creciente número de investigadores han trabajado en el análisis algorítmico y teórico de estos sistemas, creando en unos pocos años una nueva línea de investigación,

fusionando conceptos de disciplinas tan distantes como estadística, análisis funcional, optimización y maquinas de aprendizaje.

Se ha trabajado sobre la generalización del método y su extensión al caso de multclasificación por parte de Weston y Watkins; Platt, Cristianini y Shawe-Taylor; Vapnik.

Jaakkola y Haussler, Vapnik y Chapelle, Weston y Herbrich, Wahba, Lin, Zhang, Opper y Winther en sus trabajos proporcionan un análisis del error esperado de la validación cruzada y Vapnik presenta el límite del error esperado en términos del margen y el radio de la esfera más pequeña que contiene un vector de soporte [MORA05].

3.8.1. Metodología de las SVM

Las SVM son una consecuencia práctica de la teoría de aprendizaje y su estudio es útil por características como la de evitar óptimos locales para siempre buscar un óptimo global, la alta capacidad de generalización y el uso de funciones de núcleo o *kernel*. Las SVM, a diferencia de otros métodos de clasificación, tienen la ventaja de no requerir ningún tipo de hipótesis sobre la densidad de probabilidad de los rasgos, y de ser convenientes en términos de la dimensionalidad del problema, entre otros.

Considérese un conjunto T con n datos de entrenamiento n -dimensionales (\mathbf{x}_i, y_i) , con su respectiva etiqueta o clase (y_i) , tal como se presenta en (18)

$$\mathbf{x}_i \in \mathbb{R}^N \quad \text{y} \quad y_i \in \{+1, -1\} \quad (18)$$

Se busca estimar una función $f(\mathbf{x})$ tal que para una entrada en \mathbb{R}^N produzca una salida en $\{\pm 1\}$, según se presenta en (19)

$$f(\mathbf{x}): \mathbb{R}^N \rightarrow \{+1, -1\} \quad (19)$$

Se puede clasificar correctamente un nuevo dato, considerando que la función $f(\mathbf{x})$ para este nuevo dato es generada con la misma distribución de probabilidad $P(\mathbf{x}_i, y_i)$ de los datos de entrenamiento. Como no se imponen restricciones en la función que se escoge, se pueden cometer errores en la estimación, ya que aunque se realice un buen entrenamiento, no necesariamente tiene una buena generalización para datos desconocidos. Por tanto, el aprendizaje perfecto no es posible y la minimización del error de entrenamiento no implica que no haya error en la prueba.

3.8.2. Análisis del caso linealmente separable

3.8.2.1 Hiperplano clasificador óptimo

Los clasificadores con SVM se fundamentan en la obtención de hiperplanos que separen los datos de entrenamiento en dos subgrupos. Entre cada una de las clases etiquetadas como $\{-1, +1\}$, existe un único hiperplano óptimo de separación (OSH). Se busca que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano sea máxima, con la intención de forzar la generalización de la máquina de aprendizaje [BURG98] [MOJJ06].

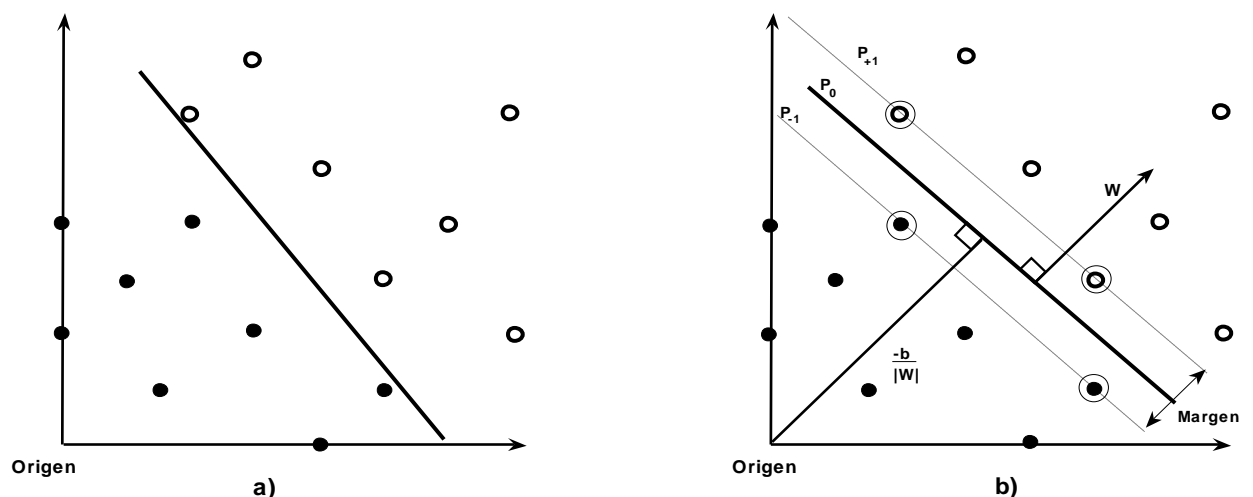


Figura 12. Hiperplanos que separan correctamente un conjunto de datos. a) Hiperplano de separación de datos. b) OSH con un mayor margen de separación entre clases.

El hiperplano óptimo de separación (OSH) de la figura 12, se presenta en (20)[†]:

$$g(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b = 0 \quad (20)$$

Para maximizar el margen, se proponen dos planos paralelos que contienen los puntos más cercanos al OSH. Si se define como $1/\|\mathbf{w}\|$ es la distancia entre cualquiera de los planos al OSH, las ecuaciones de los planos están dadas por (21) y (22)[‡]:

$$P_{+1} : (\mathbf{w} \cdot \mathbf{x}) + b = +1 \quad (21)$$

$$P_{-1} : (\mathbf{w} \cdot \mathbf{x}) + b = -1 \quad (22)$$

El margen definido como la distancia perpendicular entre (21) y (22) está dado por (23). \mathbf{x}_{+1} y \mathbf{x}_{-1} son puntos en los hiperplanos +1 y -1

$$\begin{aligned} \left[(\mathbf{w} \cdot \mathbf{x}_{+1}) + b \right] - \left[(\mathbf{w} \cdot \mathbf{x}_{-1}) + b \right] &= +1 - (-1) \\ \mathbf{w} \cdot (\mathbf{x}_{+1} - \mathbf{x}_{-1}) &= 2 \\ \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_{+1} - \mathbf{x}_{-1}) &= \frac{2}{\|\mathbf{w}\|} \end{aligned} \quad (23)$$

La distancia entre el OSH y el origen de coordenadas está dada por (24)

$$\text{Si } \mathbf{w} \cdot \mathbf{x}_i + b = 0 \quad \text{y } \mathbf{x}_i = 0 \Rightarrow \frac{\mathbf{w} \cdot \mathbf{x}_i + b}{\|\mathbf{w}\|} = \frac{b}{\|\mathbf{w}\|} \quad (24)$$

Por definición, entre los dos hiperplanos paralelos al OSH (P_{+1} y P_{-1}), no deben existir datos de entrenamiento y por tanto los datos deben cumplir con (25) y (26).

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \quad \text{para } y_i = +1 \quad (25)$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \quad \text{para } y_i = -1 \quad (26)$$

[†] Un plano N dimensional de la forma $a_n x_n + a_{n-1} x_{n-1} + \dots + a_1 x_1 + a_0 = 0$, se puede reescribir de la forma $(\mathbf{w} \cdot \mathbf{x}) + b = 0$.
 $\mathbf{w} = [a_n, a_{n-1}, \dots, a_1]$ es un vector perpendicular al hiperplano, $\mathbf{x} = [x_n, x_{n-1}, \dots, x_1]$, y $b = a_0$

[‡] La distancia perpendicular que hay entre cualquier punto a un plano es $\frac{(\mathbf{w} \cdot \mathbf{x}) + b}{\|\mathbf{w}\|}$

Luego la función decisión $f_{w,b}(x_i) \geq y_i$, corresponde al signo que resulta de evaluar un dato en la ecuación del OSH (20), tal como se presenta en (27).

$$f_{w,b}(x_i) \geq \text{sign} [w \cdot x_i + b] \quad (27)$$

Combinando (25) y (26) se obtiene (28).

$$y_i(w \cdot x + b) \geq 1 \quad (28)$$

Si existe un hiperplano que satisfaga (28), se dice que los datos son linealmente separables. Para encontrar el OSH se debe maximizar el margen (23), teniendo en cuenta la restricción (28), lo que es equivalente a resolver el problema planteado en (29) y (30).

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (29)$$

Sujeto a
$$y_i(w \cdot x_i + b) \geq 1, \forall i \quad (30)$$

La función (29) se llama función objetivo, y junto con (30) se conforma un problema de optimización cuadrático con restricciones. Los problemas de este tipo se pueden tratar introduciendo el método de multiplicadores de *Lagrange*.

3.8.2.2 Solución al problema de optimización con restricciones

Cada restricción de la forma $R_i \geq 0$, se multiplica por α_i (un multiplicador de Lagrange positivo), y se restan de la función objetivo, para así formar la *función de Lagrange* presentada en (31) [BURG98].

$$L(w, \alpha, b) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(w \cdot x_i + b) - 1] \quad (31)$$

La función de Lagrange (34) se debe minimizar con respecto a las variables primarias w y b , y maximizada sobre los α_i para encontrar el punto de silla [SCHO02]. Para el caso de $y_i(w \cdot x_i + b) - 1 > 0$, el correspondiente α_i debe ser cero, debido a que éste es el valor que maximiza a (31).

Los α_i diferentes de cero son para el caso en que $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 = 0$, que corresponden a los patrones de entrenamiento que quedan sobre los hiperplanos paralelos al OSH dados por las ecuaciones (21) y (22). Este último enunciado corresponde a las condiciones de *Karush-Kuhn-Tucker* o condiciones complementarias de optimalidad, presentadas de (32) a (34) [KUHN51].

$$\alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0, \forall i \quad (32)$$

$$\frac{\partial L(\mathbf{w}, \alpha, b)}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad (33)$$

$$\frac{\partial L(\mathbf{w}, \alpha, b)}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0 \quad (34)$$

La solución de \mathbf{w} en (33) queda en función de un subconjunto de patrones de entrenamiento cuyo multiplicador de Lagrange es diferente de cero. Es decir, el soporte de \mathbf{w} está en los patrones de entrenamiento más cercanos al OSH, es decir los vectores de soporte. De aquí el nombre de *Máquinas de Soporte Vectorial*.

Reemplazando (33 y 34) en (31), se eliminan las variables primarias \mathbf{w} y b para obtener el *problema de optimización dual de Wolfe*, el cual se resuelve mediante (35).

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (35)$$

Sujeto a :

$$\alpha_i \geq 0, \forall i \text{ y } \sum_{i=1}^n \alpha_i y_i = 0$$

$$(36)$$

Este problema puede ser resuelto con métodos de programación cuadrática estándar [BURG98]. Una vez obtenido el vector α , se pueden obtener los parámetros \mathbf{w} y b con (33) y (32) respectivamente:

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad b = y_i - \mathbf{w} \cdot \mathbf{x}_i \quad (37)$$

Ahora, la ecuación del OSH y la función de decisión se pueden expresar como se presentan en (38) y (39).

$$g(\alpha) = \sum_{i=1}^n \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i) - b \quad (38)$$

$$f(\alpha) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i) - b \right) \quad (39)$$

La utilización de los multiplicadores de Lagrange se debe a dos razones fundamentales:

- Las restricciones (30) están en función de α_i (36), que las hace más fácil de resolver.
- En la reformulación del problema, los datos de entrenamiento \mathbf{x}_i sólo aparecen en forma de productos punto entre ellos mismos (35), (38) y (39).

3.8.3. Análisis del caso linealmente no separable

3.8.3.1. Planteamiento general

La implementación de clasificadores basados en SVMs desarrolladas como se presenta en las secciones anteriores, puede tener altos errores, debido a que en la práctica no necesariamente existe un hiperplano separador, y si existe, no siempre es la mejor solución para el problema de clasificación. Cuando existen datos erróneos, ruido o alto traslapamiento de clases en los datos de entrenamiento, ésta solución puede no ser la mejor [MOJJ06].

Debido a lo anterior, se propone otra alternativa que busque el mejor hiperplano tolerando ruido en los datos de entrenamiento. Una solución aparente consiste en encontrar el hiperplano que conduzca al menor número de errores de entrenamiento, pero desafortunadamente, esto se convierte en un problema combinatorial difícil de aproximar. *Cortes y Vapnik* [CORT95], proponen un planteamiento diferente para las SVMs. [BENN92] Para permitir la posibilidad de ejemplos que violen la restricción (30), por la consideración de

variables de relajación (slack), lo que se denomina como margen débil, tal como se presenta en la figura 13. [MOJJ06]

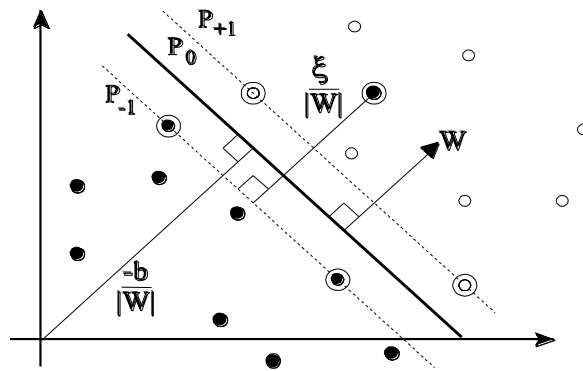


Figura 13. Hiperplano lineal clasificador para el caso no separable.

Donde:

$$\xi_i \geq 0, \forall i \quad (40)$$

La nueva restricción se presenta en (41).

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \forall i \quad (41)$$

Luego se encuentra el clasificador que mejor generaliza, controlando su capacidad de clasificación (con $\|\mathbf{w}\|$), y el límite superior del número de errores de entrenamiento $\sum_{i=1}^n \xi_i$. Una posible forma de obtener el hiperplano óptimo con margen débil es minimizando la función (42) sujeta a (43) [MOJJ06].

$$\min_{\mathbf{w}, b} \frac{1}{2} (\mathbf{w} \cdot \mathbf{w}) + C \sum_{i=1}^n \xi_i \quad (42)$$

$$\text{sujeto a: } y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \forall i \quad (43)$$

Un alto valor del parámetro C corresponde a una alta penalización a los errores. Con los multiplicadores de Lagrange el problema se transforma en (44), sujeto a (45) [MOJJ06].

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi_i \cdot \phi_j \rangle \quad (44)$$

$$\text{sujeto a: } 0 \leq \alpha_i \leq C, \forall i \text{ y } \sum_{i=1}^n \alpha_i y_i = 0$$

(45)

La solución es (33). El hiperplano separador solución se puede expresar como (38), y la función de decisión como (39) [MOJJ06].

3.8.3.2. Máquinas de soporte no lineales y funciones kernel

El principio de las SVM no lineales consiste en “mapear” o establecer una relación entre el espacio de entrada y un espacio de representación de dimensión alta, a través de una función no lineal elegida a priori, tal como se presenta en la figura 14 [BOSE92].

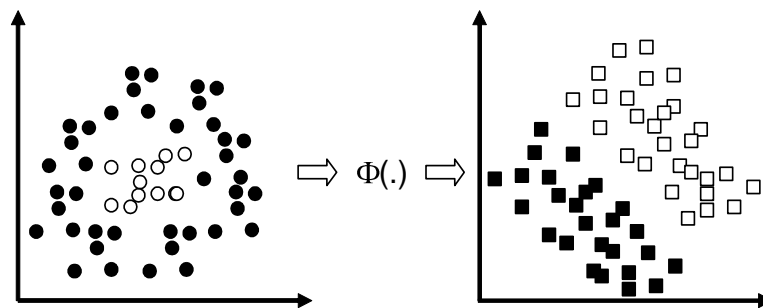


Figura 14. La SVM no lineal establece una relación del espacio de entrada con otro de representación de dimensión alta.

Por medio de una función (Φ) se trazan los datos de entrada $\phi_i \in \mathcal{R}^N$ a algún espacio de mayor dimensión y con producto punto definido (46). Este espacio se llama *espacio característico* (F) [MOJJ06].

$$\Phi: \mathcal{R}^N \rightarrow F \quad (46)$$

Así, la función (38) que depende del producto punto de los vectores en el espacio de entrada, pasa a una función que depende del producto punto de los vectores en el espacio característico, como se muestra en (47) [MOJJ06].

$$g(\mathbf{x}) = \sum_{i=1}^n \mathbf{w}_i \cdot \mathbf{y}_i = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{b}) \quad (47)$$

Entonces se define una función que sea el producto punto de los vectores en el espacio característico, presentada en (48).

$$k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \quad (48)$$

Debido a que (F) es de alta dimensión, el lado derecho de la ecuación (48) es costoso en términos computacionales. Sin embargo existe una función *kernel* (k), que puede evaluarse eficazmente y se puede demostrar que corresponde a un trazado de (Φ) en un espacio que abarca todos los productos punto [MOJJ06]. Un ejemplo para el *kernel* polinomial se muestra en (49), (50) y (51).

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}^d \text{ con } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \text{ y } d = 2 \quad (49)$$

Se tiene que:

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right]^2 \\ k(\mathbf{x}, \mathbf{y}) &= [x_1 y_1 + x_2 y_2]^2 \\ k(\mathbf{x}, \mathbf{y}) &= x_1^2 y_1^2 + 2x_1 y_1 x_2 y_2 + x_2^2 y_2^2 \end{aligned} \quad (50)$$

(50) se puede escribir como (51).

$$k(\mathbf{x}, \mathbf{y}) = \left[\begin{pmatrix} x_1^2 \\ \sqrt{2}x_1 x_2 \\ x_2^2 \end{pmatrix} \cdot \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1 y_2 \\ y_2^2 \end{pmatrix} \right] = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \quad (51)$$

Con la función *kernel* no se necesita definir explícitamente la función Φ , ya que ésta entrega directamente al resultado del producto punto, que es lo que realmente interesa para la aplicación de clasificación. Generalizando, se puede probar que por cada función *kernel* que presente una matriz definida positiva,

se puede construir una función Φ que cumpla con (48). Algunos de los *kernel* más utilizados son el polinomial (52), función de base radial (RBF) (53) y sigmoide (54).

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + c)^d \text{ para } c > 0 \quad (52)$$

$$k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} \quad (53)$$

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\mathbf{A} \mathbf{x}^T \cdot \mathbf{y} + \theta) \quad (54)$$

Incluyendo la función kernel se puede reescribir (44) sujeta a (45), como (55)

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sujeto a: } 0 \leq \alpha_i \leq C, \forall i \text{ y } \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \quad (55)$$

Las ecuaciones del OSH (38) y función decisión (39) se reescriben como (56) y (57)

$$g(\mathbf{x}) = \sum_{i=1}^n [y_i k(\mathbf{x}_i, \mathbf{x})] b \quad (56)$$

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n [y_i k(\mathbf{x}_i, \mathbf{x})] b \right) \quad (57)$$

3.8.4. Multclasificación con SVM

El planteamiento anterior sólo aborda el problema de la biclasificación (clases con valores de ± 1), pero muchos problemas son de más de dos clases ($y_i \in \{2, \dots, l\}, l > 2$). Para resolver el problema de multclasificación con máquinas de vectores de soporte se admiten dos tipos de arquitectura, las multclasificadoras y las biclasificadoras generalizadas [MOJJ06].

4. CONSTRUCCIÓN DEL MODELO COMPUTACIONAL

4.1. METODOLOGÍA

Se utilizó como metodología el Proceso de Desarrollo Unificado de Software, creada por Ivar Jacobson, Grady Booch y James Rumbaugh, miembros de Rational Software Corporation para la construcción del sistema. Este utiliza el Lenguaje Unificado de Modelado o UML (siglas en inglés), mediante este es posible establecer la serie de requerimientos y estructuras necesarias para plasmar el sistema de software previo al proceso de escribir código.

A continuación se muestra el diagrama de caso de uso:

Figura 15. Diagrama de Casos de Uso Tratamiento de Imágenes

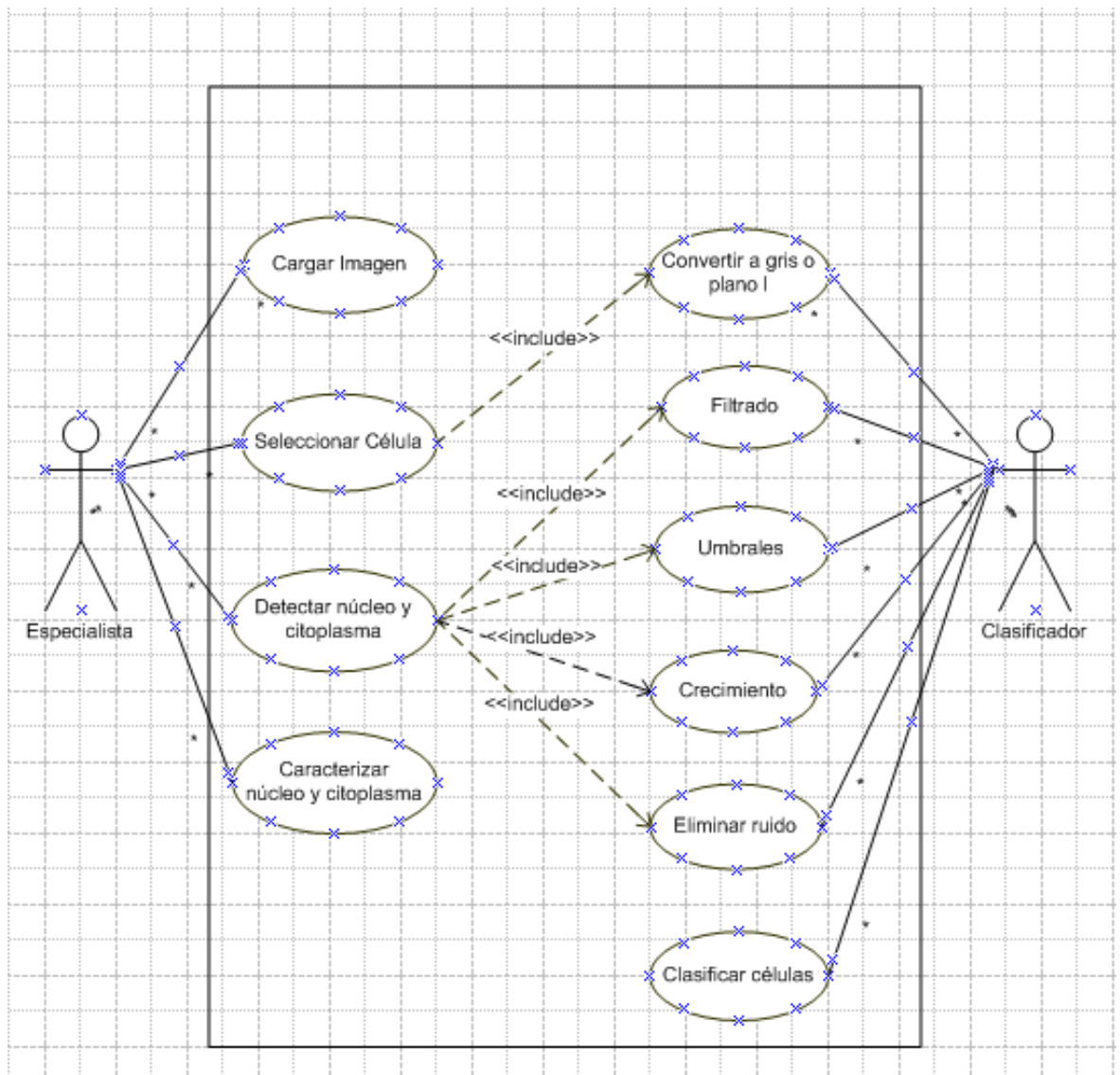


Figura 16. Diagrama de Casos de Uso, Máquinas de Soporte Vectorial

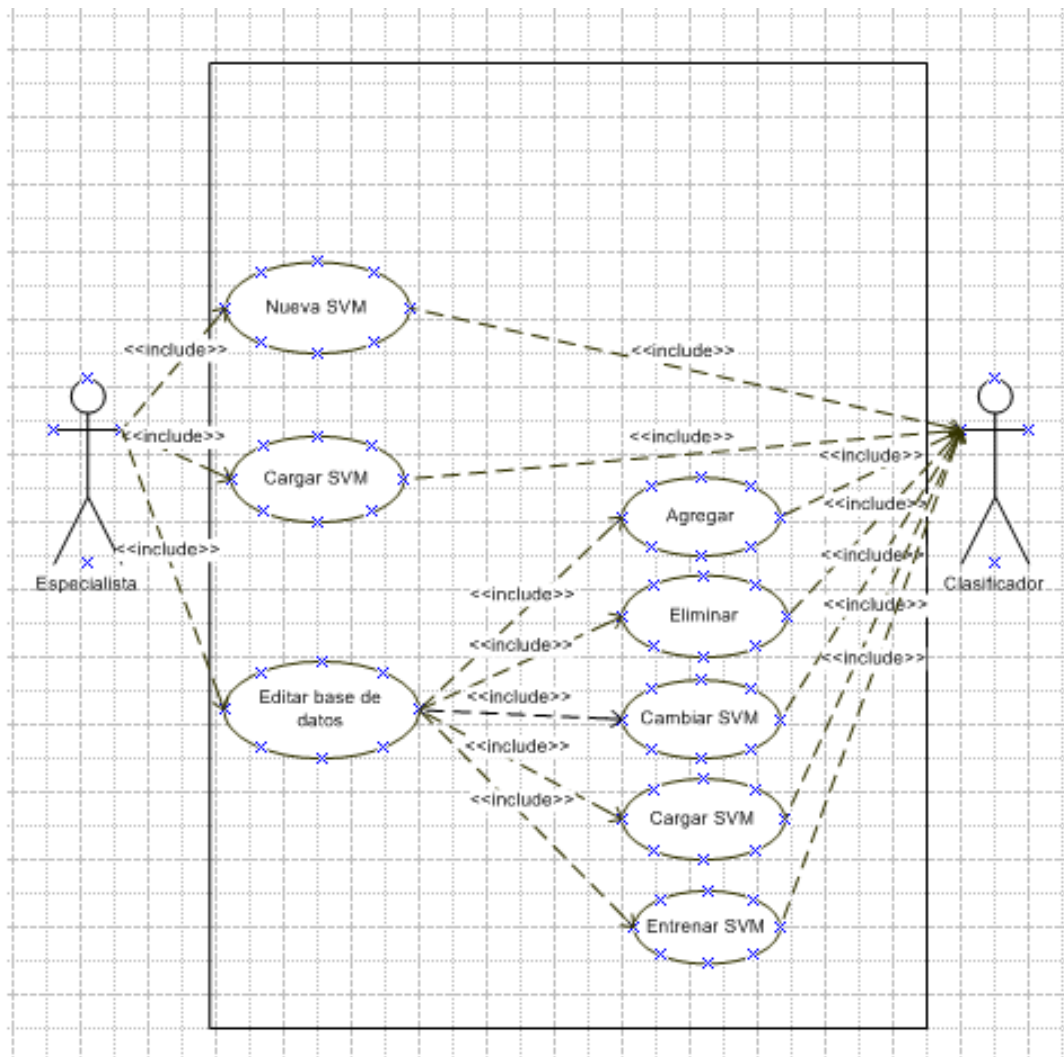


Tabla 1. Documento de descripción de los Casos de Uso

| | |
|---------------------|--|
| Nombre: | Herramienta basada en Máquinas de Soporte Vectorial para la clasificación de células escamosas de citologías cérvico uterinas. |
| Autores: | Laura Yadira Valbuena Rodríguez, Milton Antonio Caviedes Sanabria. |
| Fecha: | Enero de 2009 |
| Actores: | Especialista, Clasificador |
| Descripción: | |

Caracteriza y clasifica un campo de células escamosas obtenido en una imagen microscópica digital de una lámina de citología cérvico uterina.

Actores:

Especialista en Patología, Sistema de caracterización y clasificación(SVM).

Precondiciones:

El especialista debe contar con una lámina de citología cérvico uterina, una cámara fotográfica digital Olympus C7070 WZ, un microscopio trinocular Olympus CX31 y un acople Olympus C5060.

El zoom óptico de la cámara debe estar ajustado en 4x y deshabilitado el digital.

El objetivo a utilizar en el microscopio corresponde a 40x.

La resolución de la imagen debe ser 1200x1600 píxeles y en formato JPEG.

La aplicación debe ser ejecutada en la plataforma Matlab r2007a o superior.

Flujo Normal:

El especialista adquiere fotografías digitales de células escamosas de citologías cérvico uterinas.

El especialista carga la imagen en la aplicación.

El especialista selecciona una célula para ser analizada.

En la detección del núcleo y citoplasma se crea una interfaz para que el especialista observe los resultados de los métodos de filtrado implementados.

A continuación se procede a segmentar la imagen.

Si los algoritmos de segmentación muestran un resultado no adecuado, el especialista puede recurrir a un método manual donde determinará el valor del umbral a utilizar.

Una vez el especialista ha decidido el resultado de la segmentación el se encarga de seleccionar la parte que quiere analizar sea núcleo o citoplasma.

Una vez el especialista ha seleccionado de acuerdo a su criterio procede a la etapa de clasificación donde el sistema dará a conocer el resultado.

Flujo Alterno:

1. El especialista carga una imagen ya seleccionada.

Se continúa con los pasos posteriores.

2. Se procede a reducción del ruido con el botón trazar líneas para imágenes con bajo contraste.

3. Se utilizan los métodos de segmentación por umbralización.

Se continúa con los pasos posteriores.

Pos-Condiciones:

La célula ha sido clasificada, el especialista confrontará el resultado del sistema con su criterio y confirmará o rechazará el resultado.

4.2. RECOLECCIÓN DE MUESTRAS

Las muestras de citologías cervicouterinas eran recolectadas por los patólogos del proyecto, se tuvieron en cuenta los siguientes criterios de inclusión:

- Citologías cervicouterina con coloración de papanicolaou.
- Muestras consideradas como satisfactorias por los patólogos, cumpliendo los requisitos del sistema de Bethesda.
- Citologías normales.
- Citologías anormales.
- Citologías que presentaron cambios reactivos o lesiones preneoplásicas y neoplásicas.

El total de muestra fue de 154 imágenes de células individuales anormales y 156 datos de células benignas obtenidas por el software Diagnosis [MART04].

5. ADQUISICIÓN

Para el proceso de adquisición fueron recolectadas una serie de muestras por la Facultad de Salud de la Universidad Industrial de Santander, más específicamente por el departamento de Patología. Todas estas muestras presentan anormalidad de diferentes clases.

Para hacer el respectivo entrenamiento de la herramienta clasificadora se utilizaron los resultados obtenidos para células escamosas benignas proporcionados por el software Diagnósis [MART04].

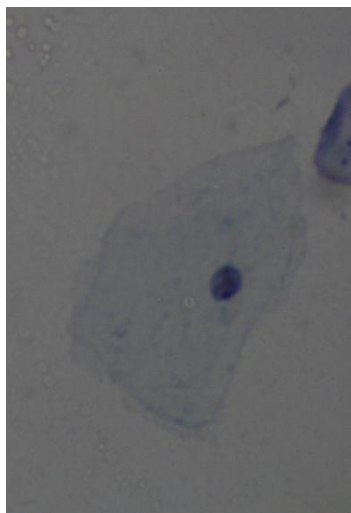


Figura 17. Imagen digital con diagnóstico de malignidad

5.1. RELACIÓN PÍXEL-MICRA

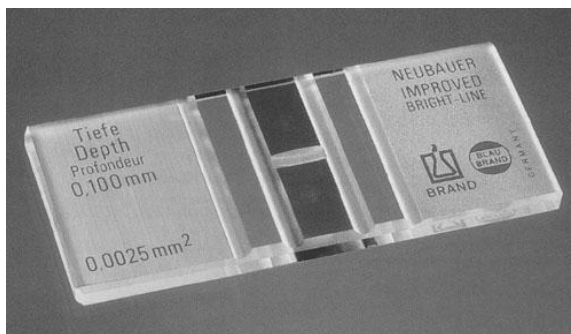
La imagen que se obtiene posee 1200 filas por 1600 columnas, pero no se contaba con una medida que permitiera conocer el tamaño del píxel, para calcular las dimensiones de las células. Con el fin de determinar la medida de

un píxel, asignadas en micras (μm), se utilizó una cámara de Neubauer de contaje celular.

Se adquirieron fotografías de los cuadros interiores, puesto que la medida de sus lados es conocida. Para la cámara de Neubauer [REMA08] utilizada, el lado del cuadro interior más pequeño tiene una medida de $62.5 \mu\text{m}$ y mediante técnicas de detección de bordes y aproximaciones estadísticas se encontró que cada lado tenía 247 píxeles; si se utilizaba: un objetivo de 40x en el microcopio, un zoom óptico de 4x para la cámara fotográfica y un lente de acople Olympus C5060.

Estas condiciones de aumento se utilizaron para todas las imágenes adquiridas. A partir de las relaciones mencionadas, se obtuvo que cada píxel tuvo un lado de $0.2530 \mu\text{m}$. En la siguiente figura se muestra la cámara de Neubauer.

Figura 18. Cámara de Neubauer.



(a) Cámara de Neubauer



(b) Cuadros interiores

6. PRE-PROCESAMIENTO

Obtenida la imagen digital se procede a la segunda etapa del tratamiento digital de imágenes. Para esta etapa se emplearon algoritmos de filtrado espacial como filtros paso bajo. También se utilizó un filtro para aumento del contraste llamado filtro morfológico, para este proceso se utilizó la conversión de imágenes en RGB a HSL, donde para el caso de la imagen digital en el modelo HSL se utilizó el plano L (luminancia).

El filtrado se realiza como la convolución de la matriz del filtro $h(x, y)$ con la imagen original $f(x, y)$, dando como resultado la imagen filtrada $g(x, y)$:

$$g(x, y) = h(x, y) * f(x, y)$$

6.1. FILTRADO

A continuación se presentan los filtros utilizados.

- **Promedio.** Este filtro promedia el valor del píxel con sus vecinos, su resultado depende del tamaño de la matriz utilizada. Un filtro de tamaño tres está definido de la siguiente manera:

$$1/9 * \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- **Promedio Circular.** Utiliza una matriz de forma circular para aplicar el promedio del píxel con sus vecinos, el resultado depende del radio. Un filtro de radio dos está definido de la siguiente manera:

$$\begin{pmatrix} 0 & 0.0170 & 0.0381 & 0.0170 & 0 \\ 0.0170 & 0.0784 & 0.0796 & 0.0784 & 0.0170 \\ 0.0381 & 0.0796 & 0.0796 & 0.0796 & 0.0381 \\ 0.0170 & 0.0784 & 0.0796 & 0.0784 & 0.0170 \\ 0 & 0.0170 & 0.0381 & 0.0170 & 0 \end{pmatrix}$$

- **Rotacional.** El resultado del filtro es la aproximación al movimiento de una cámara, que depende de la longitud y el ángulo. Para 0° y 90° funciona como un filtro promedio de una fila o una columna respectivamente. Si la longitud es tres y el ángulo de 45° se obtiene:

$$\begin{pmatrix} 0 & 0.0754 & 0.1883 \\ 0.0754 & 0.3215 & 0.0754 \\ 0.1883 & 0.0754 & 0 \end{pmatrix}$$

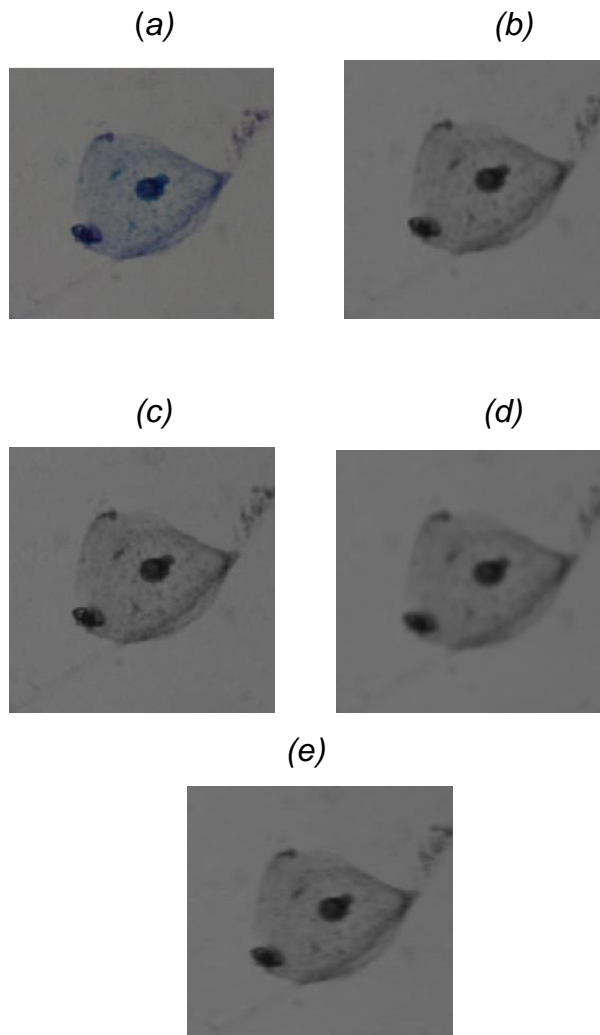
- **Gauss.** Se aplica un filtro Gaussiano rotacionalmente simétrico, el resultado depende del tamaño de la matriz y de la desviación estándar utilizada (sigma). Un filtro de tamaño tres y sigma ½ está definido de la siguiente manera:

$$\begin{pmatrix} 0.0113 & 0.0838 & 0.0113 \\ 0.0838 & 0.6193 & 0.0838 \\ 0.0113 & 0.0838 & 0.0113 \end{pmatrix}$$

La figura presenta el resultado de la aplicación de los filtros mencionados para una célula escamosa.

Figura 19. Filtro con promedio, gauss, circular y rotacional

(a)Imagen Original. (b)Filtrado Promedio. (c)Filtrado Gauss. (d)Filtrado Circular. (e)Filtrado Rotacional.



6.2. FILTRADO MORFOLÓGICO

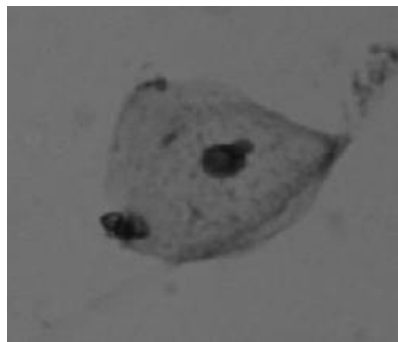
Después de cambiar el modelo de color, se realizó un filtrado basado en morfología matemática en escala de grises, con el propósito de ampliar la diferencia que existe entre los valores de los píxeles de los núcleos y los citoplasmas.

El filtro que se utilizó se construyó a partir de las operaciones morfológicas top-hat y bottom-hat, su funcionamiento está dado por los siguientes pasos:

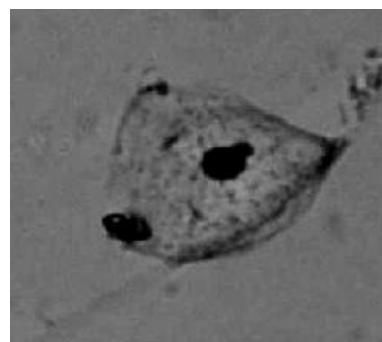
- Se determinó un elemento estructurante plano de forma circular y radio 25.
- Se aplica la operación morfológica top-hat al corte de la imagen correspondiente al plano de la luminosidad.
- Se aplica la operación morfológica bottom-hat al corte de la imagen correspondiente al plano de la luminosidad.
- Se suma la imagen de la luminosidad con la imagen resultado de la operación top-hat.
- Se resta la imagen de la suma del paso anterior con la imagen resultado de la operación bottom-hat.

La siguiente figura presenta el resultado de filtrar la imagen de la luminosidad [MART07].

Figura 20. Filtrado Morfológico



Plano de la Luminosidad



Plano de la Luminosidad Filtrado

6.3. ELIMINACIÓN DEL RUIDO Y AUMENTO DEL CONTRASTE

No todas las imágenes digitales poseen las mismas características de luminosidad y contraste, para el caso de imágenes con bajo contraste en donde no se logra diferenciar de manera clara el citoplasma con respecto al fondo debido a que este presenta bastante parecido en color con el citoplasma. Para este inconveniente se empleó un procedimiento de reducción del ruido, el cual se desarrolla de la siguiente manera.

- El especialista deberá reconocer si esta imagen digital necesita este procedimiento, luego procederá a mejorarla.
- El procedimiento consisten en trazar líneas alrededor del citoplasma, estas “borran” la imagen por donde pasan. Este “borrado” consiste simplemente en cambiar los pixeles de la imagen en escala de grises por *unos* o sea color blanco en TDI.
- Una vez trazadas las líneas se procede a aumentar el contraste de la imagen, para esto se utilizó la función *histeq*, que aumenta el rango dinámico de la imagen.

Para el paso 3 también se utilizó la función *imadjust*, pero sólo funciona para cierto tipo de imágenes tanto para segmentar el núcleo como el citoplasma.

La figura 21 muestra el proceso obtenido.

Figura 21. Eliminación del Ruido para imágenes con bajo contraste

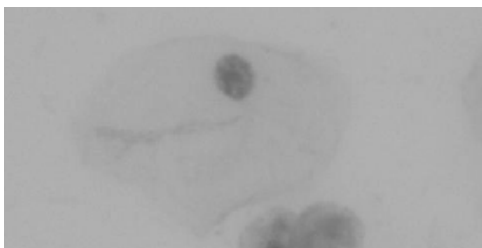
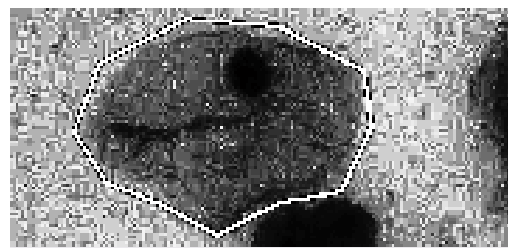


Imagen en Escala de Grises



Reducción del ruido y aumento del Contraste

La figura 22 muestra el proceso por ajuste de la intensidad. Para este caso se utiliza la función *imadjust*.

Figura 22. Ajuste de intensidad.

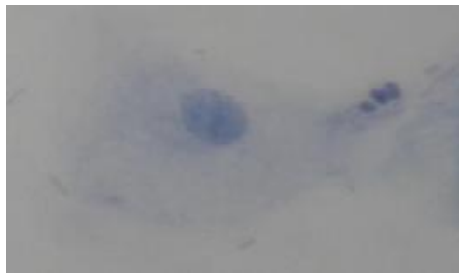
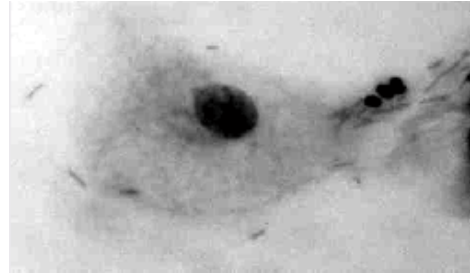
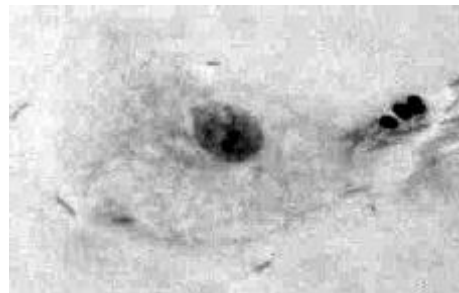


Imagen Original



Ajuste de contraste



Filtrado Morfológico

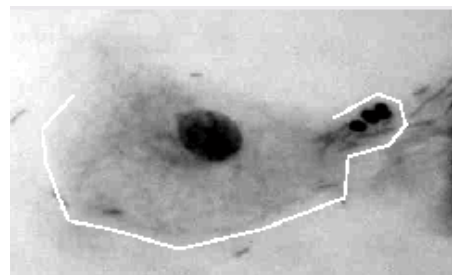
Como se puede observar, no siempre el filtrado morfológico obtiene buenos resultados, por esta razón el uso del ajuste de contraste es mejor en este caso.

También es posible utilizar la reducción del ruido en este caso.

Figura 23. Eliminación del ruido utilizando ajuste del contraste



Imagen Original



Eliminación del ruido

7. SEGMENTACIÓN

En esta etapa se procede a separar de la imagen digital el núcleo y citoplasma de la célula, para esto se utilizaron varios procedimientos como: umbralización, crecimiento de regiones y segmentación manual.

7.1. UMBRALIZACIÓN

El primer método utilizado para detectar los núcleos y citoplasma de las células escamosas consiste en calcular un umbral, basado en la separación que presenta el histograma. En este trabajo se utilizó un proceso de umbralización recursivo e iterativo que se presenta a continuación [MART07]:

- Se calcula el umbral con el método de Otsu.
- Se realiza la operación morfológica de la apertura binaria, para separar algunos objetos unidos.
- Se eliminan los objetos que se encuentren en el extremo de la imagen.
- Se calcula el área de todos los objetos encontrados y la respectiva mediana.
- Se eliminan los objetos cuya área es menor al 50% de la mediana.
- Se eliminan los objetos cuya área es mayor al 150% de la mediana.
- Se crea una máscara con los objetos eliminados por el criterio de la mediana.
- Se inicia nuevamente el proceso de umbralización con los núcleos que han sido eliminados, a partir de la máscara creada, hasta que la mediana de los objetos eliminados sea menor al doble de la mediana de los objetos no eliminados.

7.2. CRECIMIENTO DE REGIONES

En ciertos casos el método de umbralización no funciona adecuadamente debido a ciertas células presentan variaciones en la coloración nuclear, posiblemente a causa de alguna anomalía celular.

Para resolver esta situación, se recurrió a un algoritmo de segmentación implementado en la caracterización de células endocervicales, el cual está basado en el crecimiento de regiones a partir de puntos semillas [MART07].

Figura 24. Crecimiento de regiones. Segmentación del Núcleo

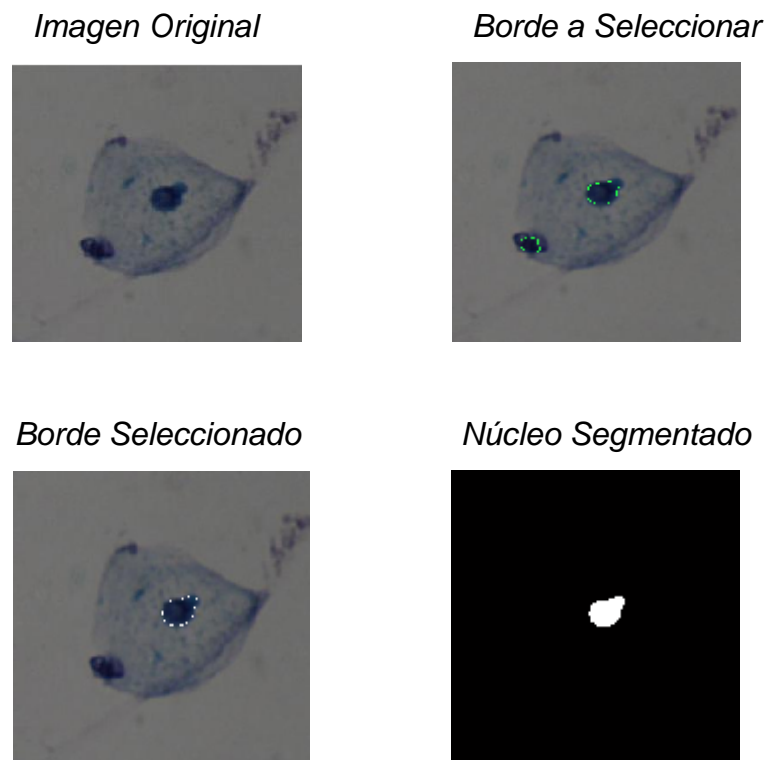
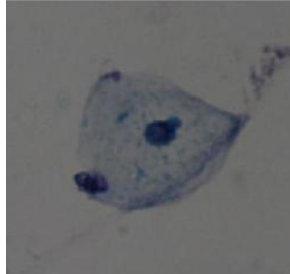
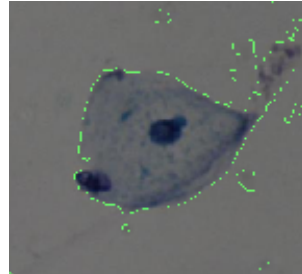


Figura 25. Crecimiento de regiones. Segmentación del Citoplasma

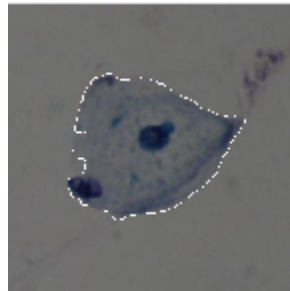
Imagen Original



Borde a Seleccionar



Borde Seleccionado



Citoplasma Segmentado



El principal inconveniente de este método es el tiempo de computación que requiere, el cual toma varios minutos.

7.3. SEGMENTACIÓN MANUAL

Consiste en localizar el umbral por experimentación. La figura presenta el resultado de este método. Los pasos son los siguientes:

- Se establece el umbral manualmente.
- Se realiza la operación morfológica de la apertura binaria.
- Se eliminan los objetos que se encuentren en el extremo de la imagen.
- Se calcula el área de todos los objetos encontrados y la respectiva mediana.

- Se eliminan los objetos cuya área es menor al 50% de la mediana.
- Se eliminan los objetos cuya área es mayor al 150% de la mediana.

Figura 26. Resultado de la Segmentación manual del Núcleo

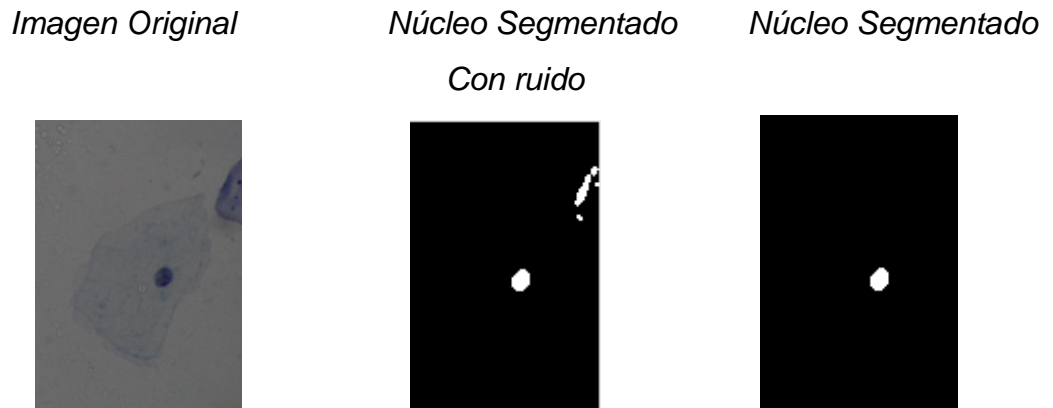
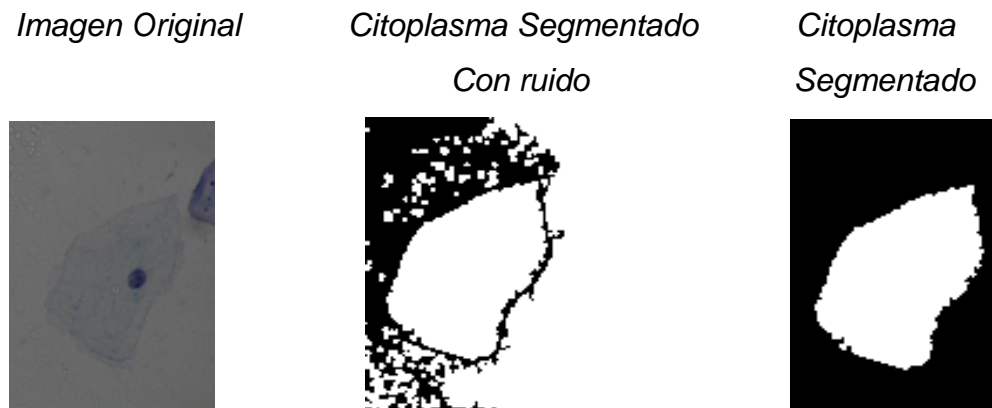


Figura 27. Resultado de la Segmentación manual del Citoplasma



Si se ha segmentado adecuadamente la imagen, se sigue con la etapa de descripción.

8. DESCRIPCIÓN

El resultado de la segmentación es un grupo de imágenes binarias, y corresponden con las regiones del núcleo, el citoplasma y sus respectivos bordes. Éstas permiten extraer los diversos parámetros de las células, utilizados para la identificación y clasificación de células escamosas de citologías cérvico uterinas.

8.1. MEDIDAS MORFOLÓGICAS

Están relacionadas con la forma y la ubicación espacial del núcleo dentro de la célula.

8.1.1. Parámetros utilizados para describir el núcleo.

- **Área:** calculada como el número de píxeles en blanco de la imagen binaria que contiene el núcleo. Para determinar la medida en micras cuadradas (μm^2) se multiplica por el factor de conversión descrito en el capítulo de adquisición.

$$\text{Área} = \sum \text{píxeles_blancos} \quad (58)$$

Donde los píxeles blancos pertenecían a un núcleo válido.

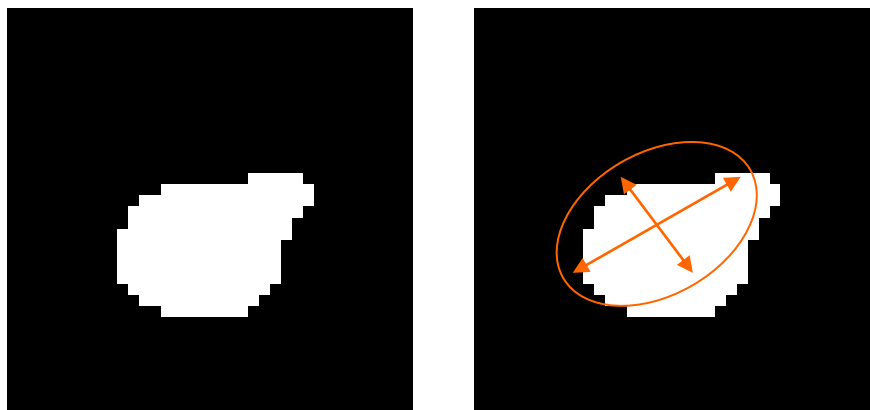
- **Perímetro:** calculado como el número de píxeles en blanco de la imagen binaria que contiene el borde del núcleo. Para determinar la medida en micras (μm) se multiplica por el factor de conversión descrito en el capítulo de adquisición.

$$\text{Perímetro} = \sum_{i=1}^n \sqrt{(x_i - b_x)^2 + (y_i - b_y)^2} \quad (59)$$

Donde: a y b son píxeles adyacentes que se encuentran en el borde y n es el número de píxeles que se encuentran en el borde.

- **Ejes:** calculados como los ejes de la elipse con el segundo momento igual a la región del núcleo.
- **Circularidad:** relación entre el cuadrado del perímetro y el área, esta medida permite conocer la tendencia circular del núcleo. Si un objeto presenta esta forma geométrica el valor es 4π .
- **Diámetro mayor y menor.** Longitud de los ejes mayor y menor de la elipse que tiene el segundo momento central igual a la región del núcleo. Para determinar la medida en micras (μm) se multiplica por el factor de conversión descrito en el capítulo de adquisición.

Figura 28. Núcleo detectado y diámetro calculado

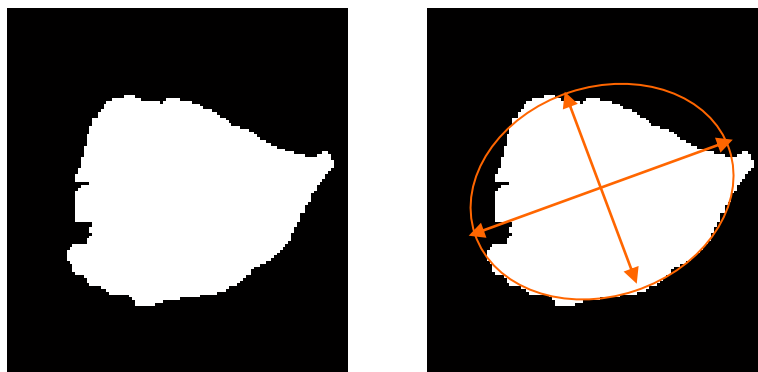


- **Irregularidad Nuclear.** Este parámetro es determinado por la varianza de la firma del núcleo.

8.1.2. Parámetros utilizados para describir el citoplasma.

- **Área:** calculada como el número de píxeles en blanco de la imagen binaria que contiene el citoplasma. Para determinar la medida en micras
- **Perímetro:** calculado como el número de píxeles en blanco de la imagen binaria que contiene el borde del citoplasma.
- **Ejes:** calculados como la longitud de los ejes de la elipse con el segundo momento igual a la región del citoplasma.
- **Diámetro mayor y menor.** Longitud de los ejes mayor y menor de la elipse que tiene el segundo momento central igual a la región del citoplasma. Para determinar la medida en micras (μm) se multiplica por el factor de conversión descrito en el capítulo de adquisición.

Figura 29. Citoplasma detectado y diámetro calculado



- **Relaciones Núcleo/Citoplasma:** llamada también Radio N/C, indica la proporción entre el tamaño del núcleo y el citoplasma; se calcula para el área, el perímetro y los ejes.

8.2. MEDIDAS DE TEXTURA

En ésta categoría se agrupan las características relacionadas con la coloración del campo celular.

- **Hipercromatismo.** Consiste en la probabilidad acumulada de los niveles de los núcleos en el intervalo $[0, 50]$ en el plano de la luminancia debida a la tinción nuclear, y está definido por la siguiente ecuación:

$$\text{hipercromatismo} = \sum_{i=0}^{50} p_i \quad (60)$$

Donde p_i es la probabilidad del nivel $i \in [0, 255]$ en el histograma

- **Pérdida de polaridad.** Esta situación era provocada por cambios en la textura del plano Matiz, para lo cual se decide utilizar los descriptores de la matriz de co-ocurrencia en éste plano, explicados en el marco teórico. Se adicionaron la entropía y la desviación estándar de los niveles en el plano.

En la descripción nuclear no es utilizado el plano del Matiz, ya que los núcleos muestran el mismo color en toda su superficie debido a la tinción usada para efectos de la lámina.

9. INTERPRETACIÓN Y RECONOCIMIENTO

En esta etapa se clasifican las células escamosas como benignas o malignas según parámetros almacenados en la base de datos para estos grupos celulares.

La base de datos de las células escamosas benignas es obtenida por el software Diagnosis [MART04], en tanto para las células malignas se hizo la respectiva base de datos, correspondiente a características de anormalidad establecidas.

Para la clasificación se ha implementado la herramienta basada en máquinas de soporte vectorial. Se usó la toolbox de Máquinas de Soporte Vectorial diseñada por el Ingeniero Electricista Germán Andrés Morales España.

La ventaja de las SVM es que poseen sólo una arquitectura para su manejo y no es necesario especificar arquitectura alguna.

La base de datos es un archivo de Matlab 2007a, a continuación se muestran las variables manejadas por la base de datos.

Datos_X, corresponde a las características de las células, área, perímetro, diámetro mayor, diámetro menor y circularidad.

Datos_Y, corresponde a las etiquetas para la clasificación de las células que en este caso es benignas o malignas.

Datos_imágenes, corresponde a los datos dados por la imagen.

Datos_nom_imag, corresponde al nombre de la imagen.

9.1. Selección de parámetros SVM

Los parámetros de la función kernel y el de penalización varían dependiendo del problema que se esté atacando. Mediante los métodos de búsqueda en malla y validación cruzada es posible obtener buenos resultados. Estos métodos son empleados por sus grandes ventajas como fácil implementación, evita el sobre-entrenamiento (ver Figura 29, obsérvese que el sobre-entrenamiento conlleva a una mala generalización de la SVM) y su alta competitividad con respecto a otros métodos de selección de hiperparámetros. Durante este proceso, se obtienen unos errores de entrenamiento y de validación. Los parámetros seleccionados son los que presenten el menor error de validación.

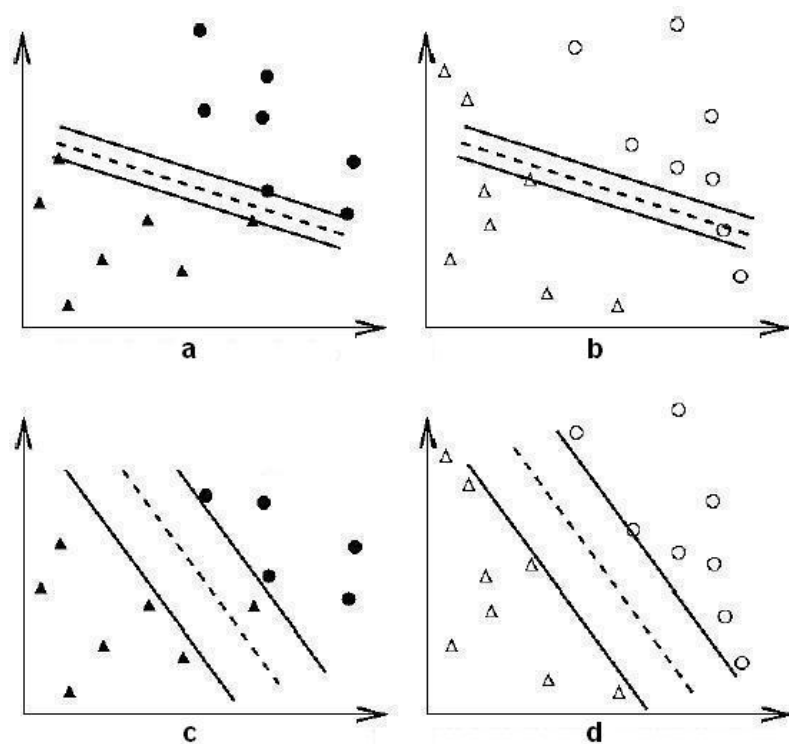


Figura 30. Representación de las SVM. a. SVM Sobre-entrenada y datos de entrenamiento. b. SVM Sobre-entrenada y datos de prueba. c. SVM sin Sobre-

entrenamiento y datos de entrenamiento. d. SVM sin Sobre-entrenamiento y datos de prueba.

9.1.1. Validación Cruzada

El método consiste en dividir el bloque de datos de entrenamiento en n partes iguales. A continuación, para un cierto valor de hiperparámetros, se utilizan $n-1$ de las n partes para entrenar la máquina y la parte restante para hallar el error de validación. Este proceso se realiza por tanto n veces, lo que permite usar todas las muestras para hallar un error de validación. Por último, se deben promediar los n valores de error de validación encontrados para obtener un solo error de validación asociado a los parámetros en uso.

9.1.2. Búsqueda en Malla

La búsqueda en malla aunque no ha sido presentada formalmente, en los últimos años los investigadores de las máquinas de aprendizaje la han empleado ampliamente por su simplicidad. Es básicamente el entrenamiento de la SVM con diferentes valores para los parámetros del kernel y de penalización. En el caso del kernel RBF, la malla es de dos dimensiones (C, σ) y se recomienda hacerla con un paso en potencias de dos, para lograr una búsqueda rápida, luego se hace una rejilla más pequeña al rededor de los parámetros que mayor precisión hayan arrojado.

10. RESULTADOS

10.1. CLASIFICACIÓN DE CÉLULAS

Para la clasificación de las células escamosas en benignas o malignas según características de normalidad y anormalidad del núcleo y citoplasma, se utilizó la herramienta de máquinas de soporte vectorial. Gracias a esta herramienta se logró hacer pruebas rápidas y confiables.

Una vez llena la base de datos con 156 células escamosas benignas y 154 células escamosas malignas, se procedió al entrenamiento de la máquina y posterior pruebas. Para este fin se crearon 10 máquinas de soporte vectorial con diferentes cantidad de características. La siguiente tabla muestra el resultado.

Características Núcleo: área, perímetro, diámetro mayor, diámetro menor y circularidad.

Características Citoplasma: área, perímetro, diámetro mayor y diámetro menor.

Características Relación Núcleo y Citoplasma: relación área, relación perímetro, relación diámetro mayor, relación diámetro menor entre el núcleo y el citoplasma.

Total de veces de entrenamiento: tres veces para cada máquina y se obtuvo promedio. Se usó el 80% de los datos para entrenamiento y 20% de los datos para pruebas tanto de células normales como anormales.

Tabla 2. Resultados de las SVM

| Máquina de soporte vectorial | Características | Porcentaje de Exactitud de Entrenamiento Promedio | Porcentaje de exactitud de Prueba Promedio |
|------------------------------|---|---|--|
| Prueba2 | Relación núcleo citoplasma y circularidad. | 100% | 97.8494% |
| Prueba3 | Núcleo. | 100% | 93.5484% |
| Prueba4 | Citoplasma, Núcleo y Relación núcleo citoplasma | 99.7312% | 91.9355% |
| Prueba5 | Núcleo y Citoplasma | 99.0592% | 92.4731% |
| Prueba6 | Núcleo, Citoplasma y Relación núcleo sólo área y perímetro. | 99.1902% | 91.3978% |
| Prueba7 | Relación núcleo citoplasma | 86.0215% | 62.3628% |
| Prueba8 | Citoplasma | 99.4633% | 91.9355% |
| Prueba9 | Citoplasma y Relación núcleo citoplasma | 99.1935% | 91.3978% |
| Prueba10 | Citoplasma y | | |

| | | | |
|--|--------------|----------|----------|
| | circularidad | 99.1935% | 93.0107% |
|--|--------------|----------|----------|

Como se puede observar en la tabla 1, la máquina llamada Prueba7 posee un porcentaje de exactitud de entrenamiento del 86.0215% y un porcentaje de exactitud de prueba del 62.3628%, esto quiere decir que al escoger estas características se puede notar que estas no son las más importantes debido al poco porcentaje como de entrenamiento y prueba.

10.2. COMPARACIÓN RED NEURONAL SOFTWARE DIAGNOSIS VERSUS MAQUINAS DE SOPORTE VECTORIAL (SVM).

Para esta comparación se tuvo en cuenta la base de datos del software Diagnósis [MART04] en donde las formas del núcleo son clasificadas como regulares e irregulares, además es el único parámetro que el software Diagnósis clasifica con redes neuronales.

La variable de salida es:

MÓDULOS: Módulos de los Descriptores de Fourier.

FORMAS: Formas de los núcleos (1=>Regular, -1=>Irregular).

CIRCULARIDADES: Circularidades de los núcleos.

A continuación se muestra una tabla con el número de datos.

Tabla 3. Datos Software Diagnósis

| Núcleos de las Células Escamosas | Total de Datos |
|----------------------------------|----------------|
| Núcleos | 436 |
| Regulares | 380 |
| Irregulares | 56 |

Tabla 4. Datos de prueba y entrenamiento para las SVM y Red Neuronal

| | |
|------------------------|----------------|
| SVM y Red Neuronal | Total de Datos |
| Datos de Entrenamiento | 349 |
| Datos de Prueba | 87 |

Se desordenaron y se promediaron los datos de prueba y entrenamiento con el fin de que la clasificación fuera aceptable. Debido al poco número de núcleos irregulares y gran número de núcleos regulares se corrió el programa varias veces para que fuese equitativo. Se modificó el programa de la red neuronal del software Diagnósis para obtener porcentajes de acierto y lograr la comparación con las SVMs. Estos fueron los resultados:

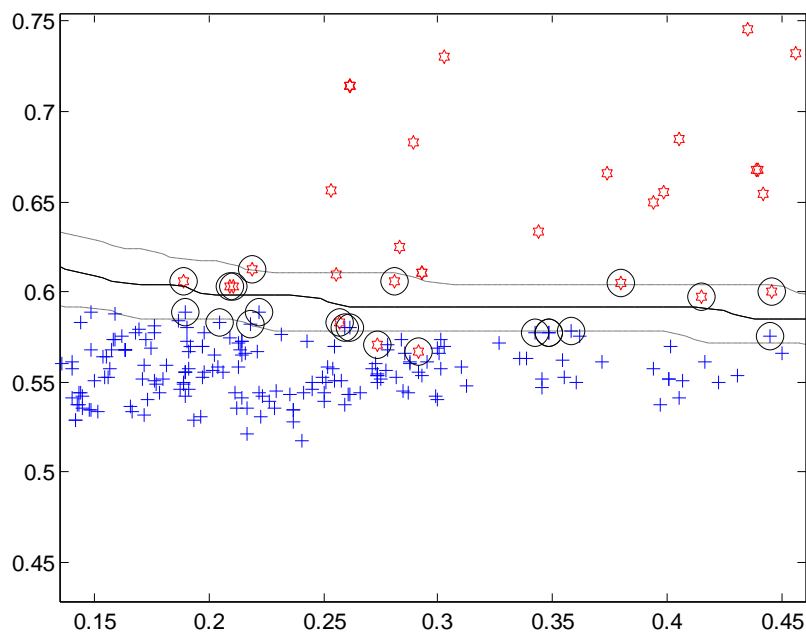
Acierto Red Neuronal

95.8966%

Acierto SVM

97.9885%

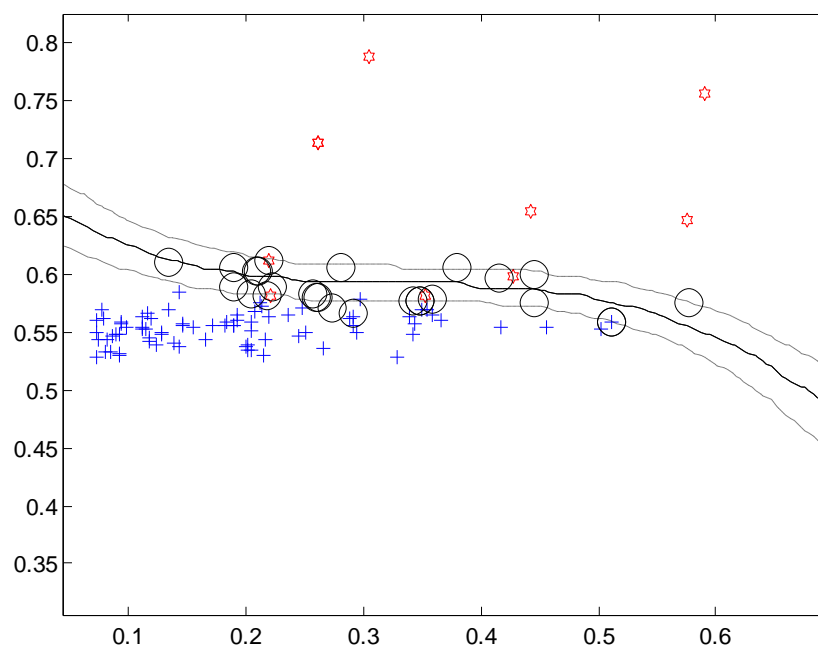
Figura 31. Datos de entrenamiento SVM



Los signos (*) equivalen a los datos de entrenamiento SVM para núcleos irregulares.

Los signos (+) equivalen a los datos de entrenamiento SVM para núcleos regulares.

Figura 32. Datos de prueba SVM



Los signos (*) equivalen a los datos de prueba SVM para núcleos irregulares.

Los signos (+) equivalen a los datos de prueba SVM para núcleos regulares.

CONCLUSIONES Y RECOMENDACIONES

Se logra demostrar que es posible una buena clasificación de las células utilizando las SVMs lo cual es el interés principal de este estudio.

Los resultados en las pruebas fueron superiores al 90% para el caso en donde se incluyen características tanto del núcleo como del citoplasma. Lo contrario sucede donde sólo se incluyen las relaciones del núcleo citoplasma y no se incluyen ninguna de las citadas anteriormente, obteniendo malos resultados como es el caso de un promedio de 86% de exactitud de entrenamiento y un promedio de 62% de exactitud de prueba, esto quiere decir que se deben escoger muy bien el conjunto de características de tal forma que se obtengan buenos resultados. Cabe resaltar que los datos de entrenamiento siempre fueron mayores que los datos de prueba.

Trabajar con la herramienta de clasificación como las máquinas de soporte vectorial es muy sencillo y rápido, comparado con las redes neuronales debido a que las máquinas de soporte vectorial no necesita ninguna arquitectura en especial, simplemente tener en cuenta los parámetros de clasificación.

Se había escrito en la justificación del anteproyecto que se iban a caracterizar células en grupo pero no fue posible debido a que la escuela de patología no contaba con la cámara para la adquisición de imágenes. Además las imágenes de las células obtenidas son todas malignas y no hubo punto de comparación con imágenes de células benignas. Para este caso se usó la base de datos del software Diagnosis como punto de comparación y por tal razón no se tuvieron en cuenta más características por la restricción de descriptores del software Diagnosis.

El campo de Inteligencia Artificial es un campo bastante amplio sobre el cual explorar y experimentar, en este caso el uso de las máquinas permitió el conocimiento de una nueva herramienta de inteligencia artificial, en este caso las Máquinas de Soporte Vectorial.

Si bien se alcanzaron los resultados obtenidos en trabajos futuros se espera clasificar células en grupo y con nuevos descriptores.

REFERENCIAS

- [BENN92] K. P. Bennett and O. L. Mangasarian (1992). Robust linear programming discrimination of two linearly inseparable set. *Optimization Methods and Software*, 1:23–34.
- [BIOP07] BIOPUR S.R.L. web side. Fecha de consulta: Enero de 2009 <http://www.biopur.com.ar/hmpapsp.htm>
- [BOSE92] B. Boser, I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifier. *Proc. 5th ACM Workshop on Computational Learning Theory*, pages: 144–152.
- [BURG98] BURGES, Christopher (1998). A tutorial on support vector machines for pattern recognition. *Data Mining And Knowledge Discovery*, 2:121–167.
- [CORT95] C. Cortes and V. Vapnik (1995). Support vector networks. *Machine Learning*, 20:273–297.
- [DECH09] DELGADO CHINCHILLA, Diana María. Prevenir el cáncer de cuello uterino es posible. En *Salud y Bienestar* No 37-38. (Enero 2009). p 1-2. http://www.observatorio.saludsantander.gov.co/publicaciones/descargar.php?id=Revista_02-03-2006-Art05.pdf&ruta=http://www.observatorio.saludsantander.gov.co/publicaciones
- [DISO05] SOLOMON Diane, MD, Nayar Ritu, MD, “El sistema Bethesda para informar la citología cervical” Ediciones Journal, Buenos Aires, Argentina, 2005.
- [FEYE01] SOSSA, M. Yesid F: “Caracterización de células del epitelio escamoso del cérvix”. <http://pci.unalmz.edu.co/Tesis/yesidsossa.pdf>. 2001
- [GOWO96] GONZALES, Rafael C. y WOODS Richard E. *Tratamiento Digital de Imágenes*. Wilmington: Addison-Wesley Iberoamericana, 1996.
- [KUAH94] KUAH K., M. Bodruzzaman, and S. Zein-Sabatto, “A Neural Network-Based Text Independent Voice Recognition System”. Southeastcon '94. ‘Creative Technology Transfer – A Global Affair’. Proceedings of the 1994 IEEE. Department of Electrical and Computer Engineering. Tennessee State University. Nashville, TN 37209-1561. 1994.
- [KUHN51] H. W. Kuhn and A. W. Tucker (1951). Nonlinear programming. *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pages: 481–492. University of California Press.
- [MART04] MARTÍNEZ, A. Víctor E. “Implementación de un modelo computacional para clasificación normal-displásica de las células escamosas de citologías cérvico uterinas”. Trabajo para optar por el título de Ingeniero de Sistemas. Universidad Industrial de Santander 2004.
- [MART07] MARTÍNEZ, A. Víctor E. “Modelo Computacional para Caracterización de Células Endocervicales”. Trabajo de investigación para optar por el título de Master en Informática. Universidad Industrial de Santander 2007.
- [MOJJ06] MORA, Juan José, Métodos de Clasificación Basados en el Conocimiento- MCBC, Tesis Doctoral, Universidad de Girona, Cataluña-España, 2006.
- [MORA05] MORALES Germán, GÓMEZ, Álvaro “Estudio e implementación de una herramienta basada en máquinas de soporte vectorial aplicada a la localización de fallas en sistemas de distribución”. Trabajo para optar por el título de Ingeniero Electricista. Universidad Industrial de Santander. 2005.
- [REMA08] REINA, Manuel. Técnicas de contaje celular. Universidad de Barcelona. Documento en línea. Fecha de consulta: Noviembre de 2008. <http://www.ub.es/biocel/wbc/tecnicas/contajecelular.htm>
- [SCHO02] B. Scholkopf, A. Smola, “Learning with kernels support vector machines, Regularization, Optimization and Beyond”. The MIT Press, Cambridge. 2002
- [VIEF08] ENCICLOPEDIA, Virtual. Cáncer Cervical. En Ferato. (Noviembre 2008). http://www.ferato.com/wiki/index.php/C%C3%A1ncer_Cervical
- [WIKI08] ENCICLOPEDIA, Virtual. Fecha de Consulta: Enero de 2009 http://es.wikipedia.org/wiki/Prueba_de_Papanicolaou

ANEXOS

A. FUNCIONAMIENTO DE LA HERRAMIENTA SOFTWARE

La herramienta software se desarrolló totalmente en Matlab R2007a. El primer paso para ejecutar la aplicación es la instalación de la toolbox de máquinas de soporte vectorial, para esto se copia el archivo svm_tool a la carpeta de donde está guardado el Matlab por ejemplo c:/archivos de programa/Matlab R2007a/toolbox, en la carpeta toolbox se copia la toolbox de SVM. Luego se inicia el Matlab, se da en archivo y luego en Set Path como se muestra en la figura 32.

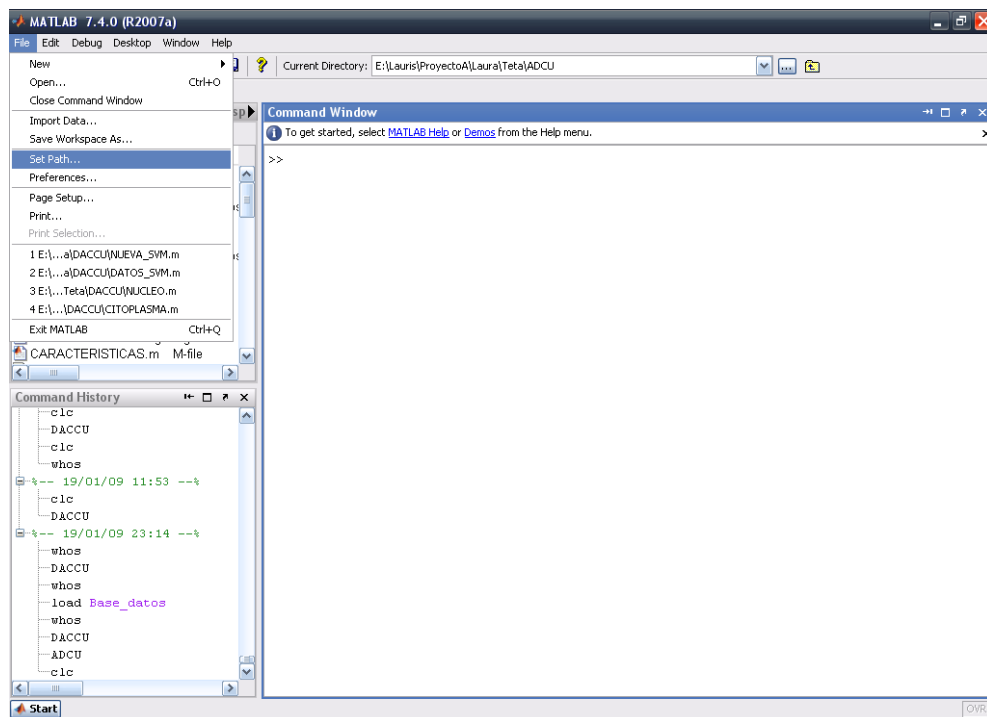


Figura 33. Instalación de la toolbox SVM

Luego se despliega una ventana, en esta ventana se oprime el botón Add Folder y se busca la toolbox ya guardada en la carpeta toolbox y le se le oprime save.

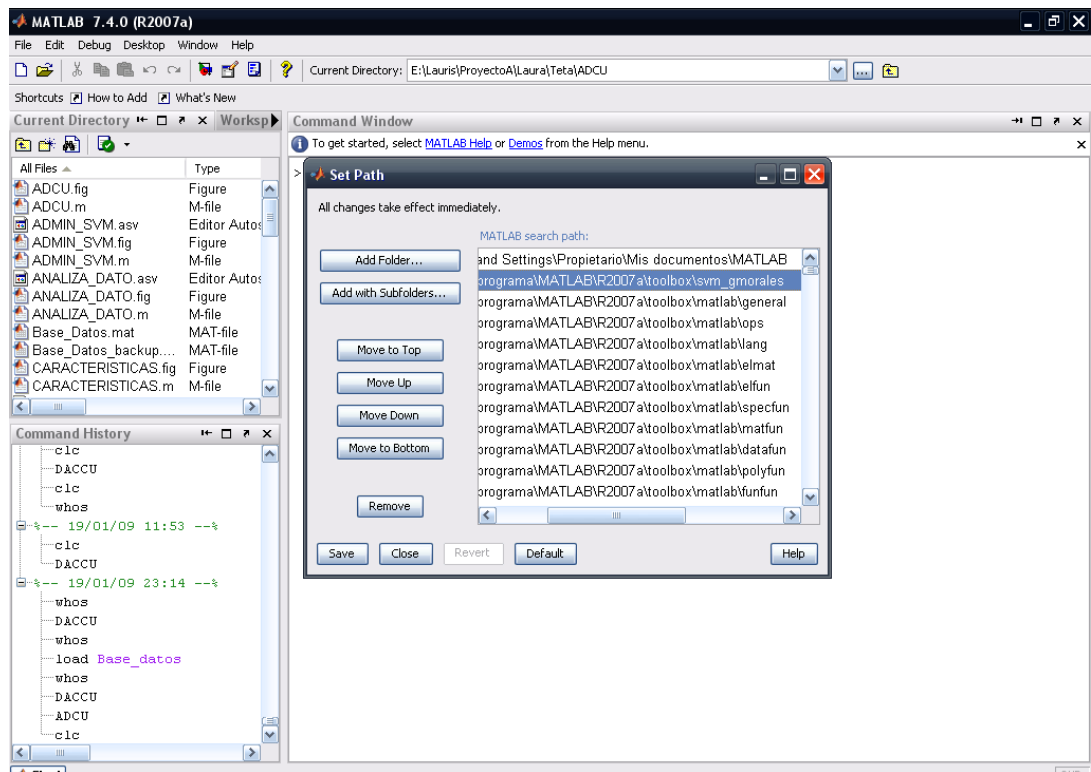


Figura 34. Salvar toolbox SVM

Una vez realizados estos pasos se procede a abrir la herramienta software escribiendo el comando para abrir la interfaz gráfica.

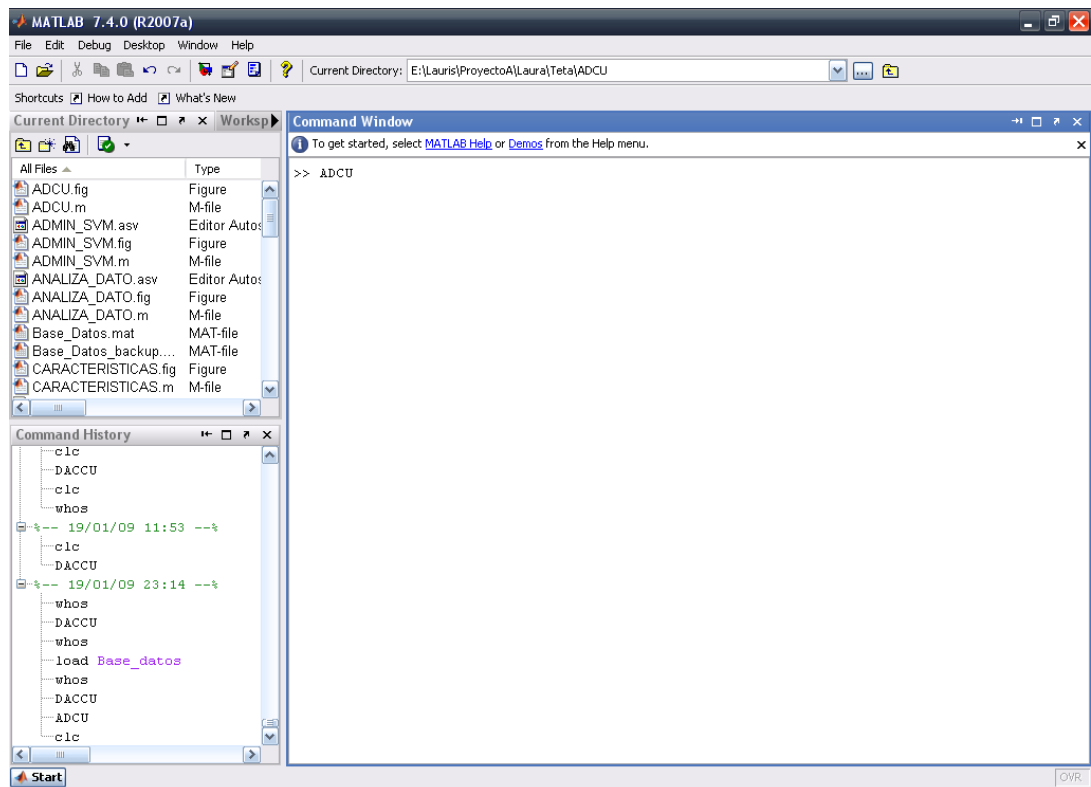


Figura 35. Inicio de la Herramienta Software

Luego de ejecutar en la ventana de comandos la línea ADCU, se despliega la ventana principal, que habilitará las opciones de procesamiento y configuración. Se encuentran las opciones disponibles: Analizar Célula, Configurar, Acerca de y Ayuda.

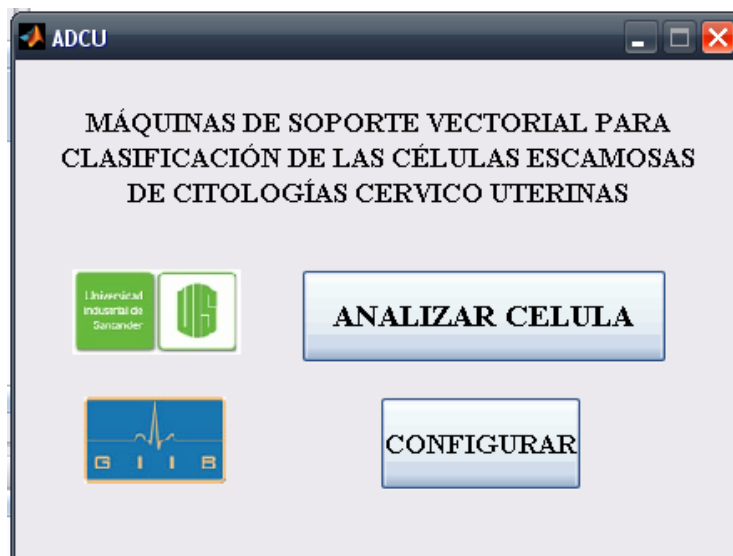


Figura 36. Ventana Principal de la herramienta software

Al utilizar la opción ANALIZAR CÉLULA se despliega una ventana de exploración con el propósito de buscar la imagen celular que se desea analizar. El software sólo debe ser utilizado para el análisis de imágenes de células escamosas cérvico uterinas. La figura muestra la forma en que se selecciona una imagen celular.

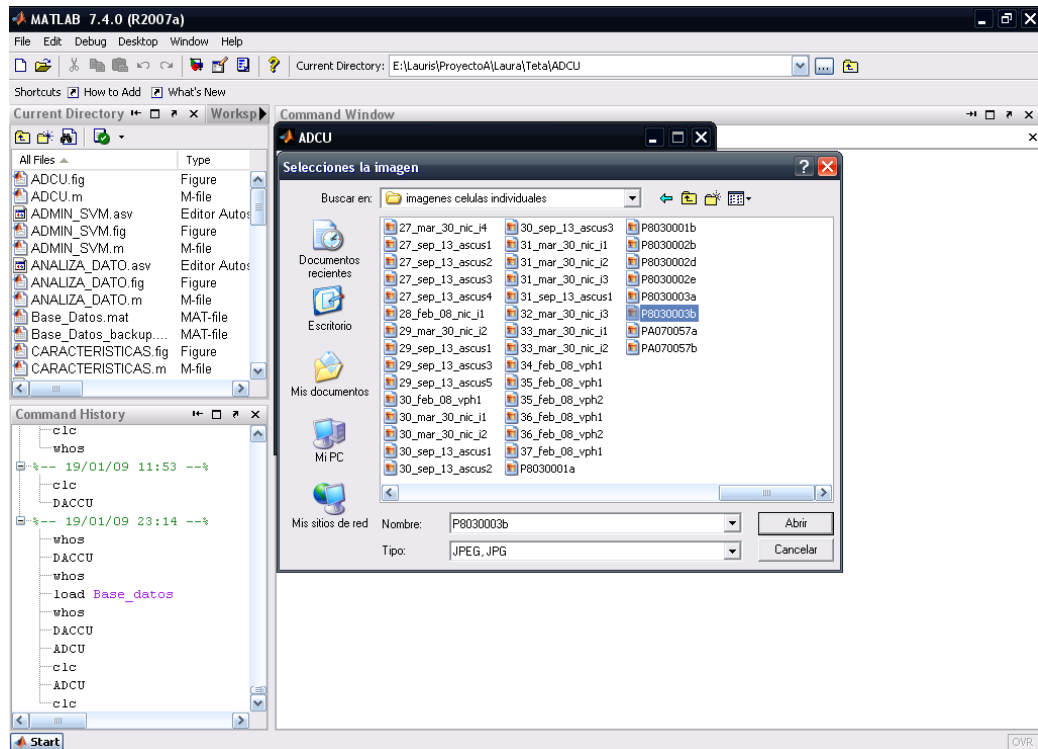


Figura 37. Opción abrir imagen

Al abrir la imagen se dan estas opciones, como se muestra en la figura



Figura 38. Opción seleccionar

Una vez seleccionada la imagen se procede a analizar núcleo o analizar citoplasma. La siguiente figura muestra el análisis para el núcleo

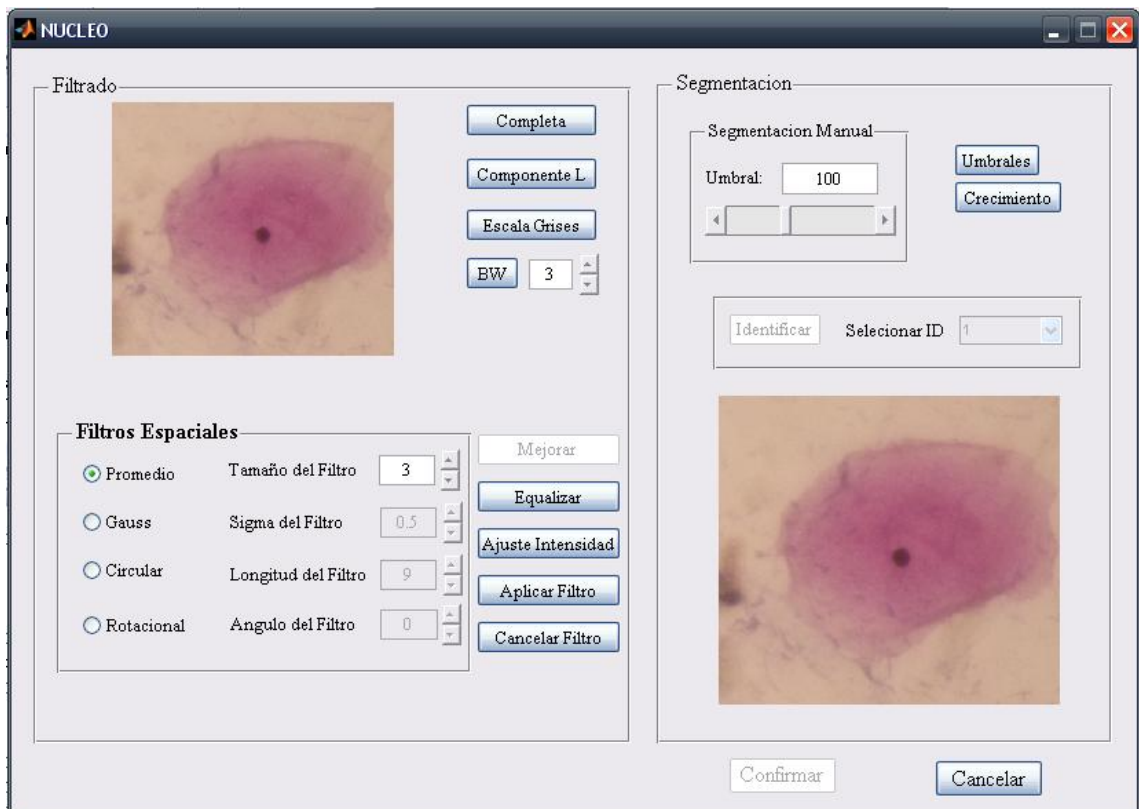


Figura 39. Ventana para el análisis del núcleo

La ventana presenta dos áreas para la imagen, en la primera se presenta todo el pre-procesado que se le hace a la imagen y en la segunda se muestra la segmentación. Ahora se procede al filtrado y al cambio de intensidad de la imagen utilizando los diferentes planos de color en RGB y HSL, para este caso se usa el plano L. Se muestra la figura 40.

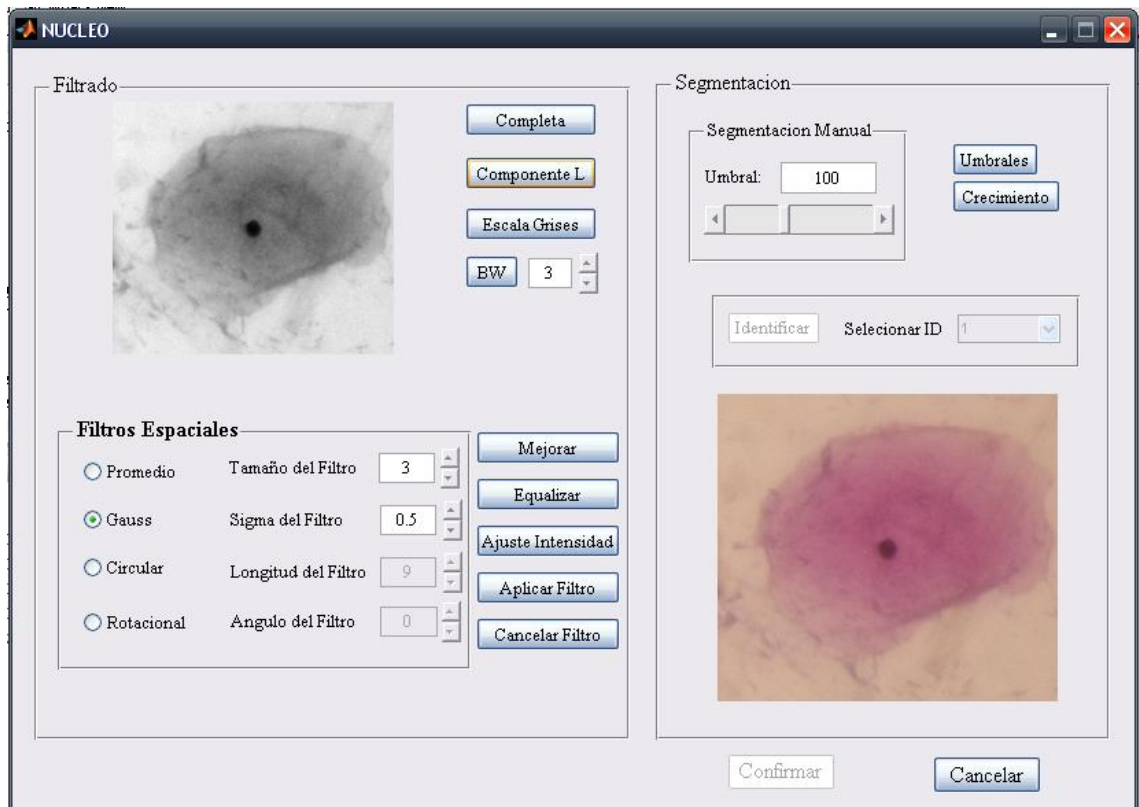


Figura 40. Opciones de pre-procesamiento

Ahora se procede al aumento del contraste con los botones Mejorar, Ecuilizar y Ajuste Intensidad. La figura muestra lo obtenido por el botón Mejorar.

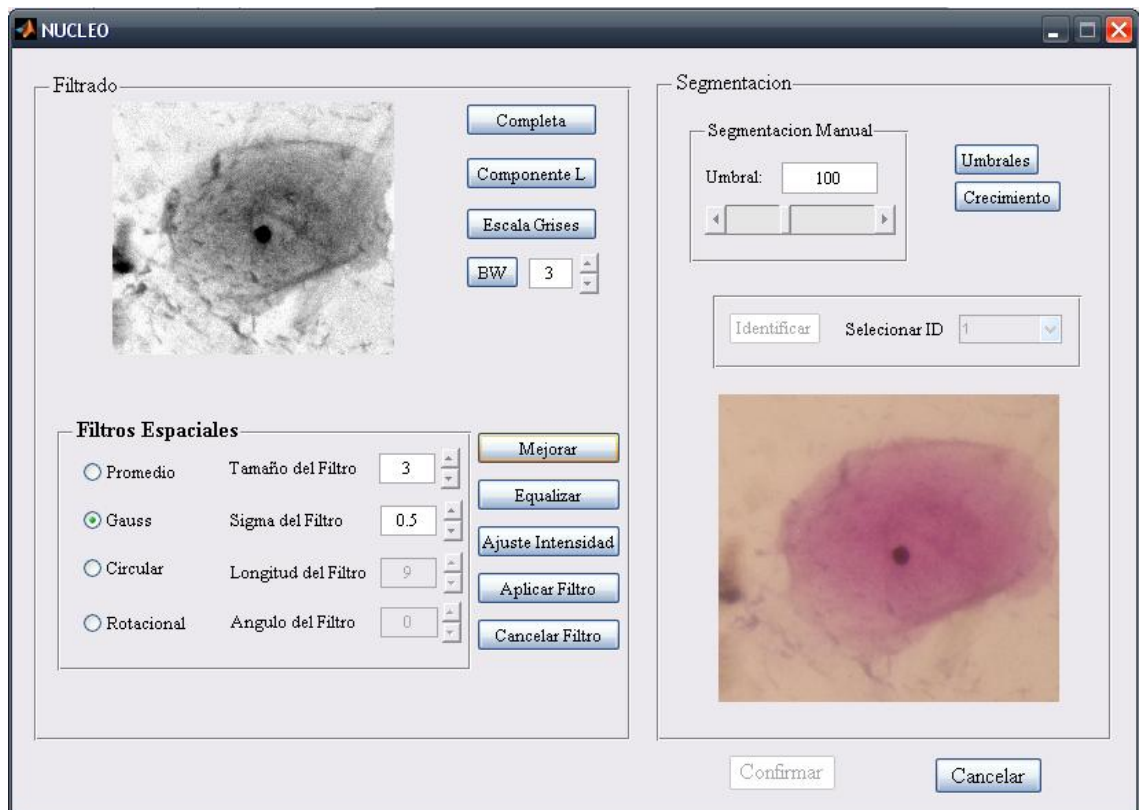


Figura 41. Opciones aumento del contraste 1.

Para que funcione correctamente el botón Mejorar la imagen debe ser pasada al plano L del modelo de color HSL.
 A continuación se muestra el resultado obtenido con el botón Equalizar

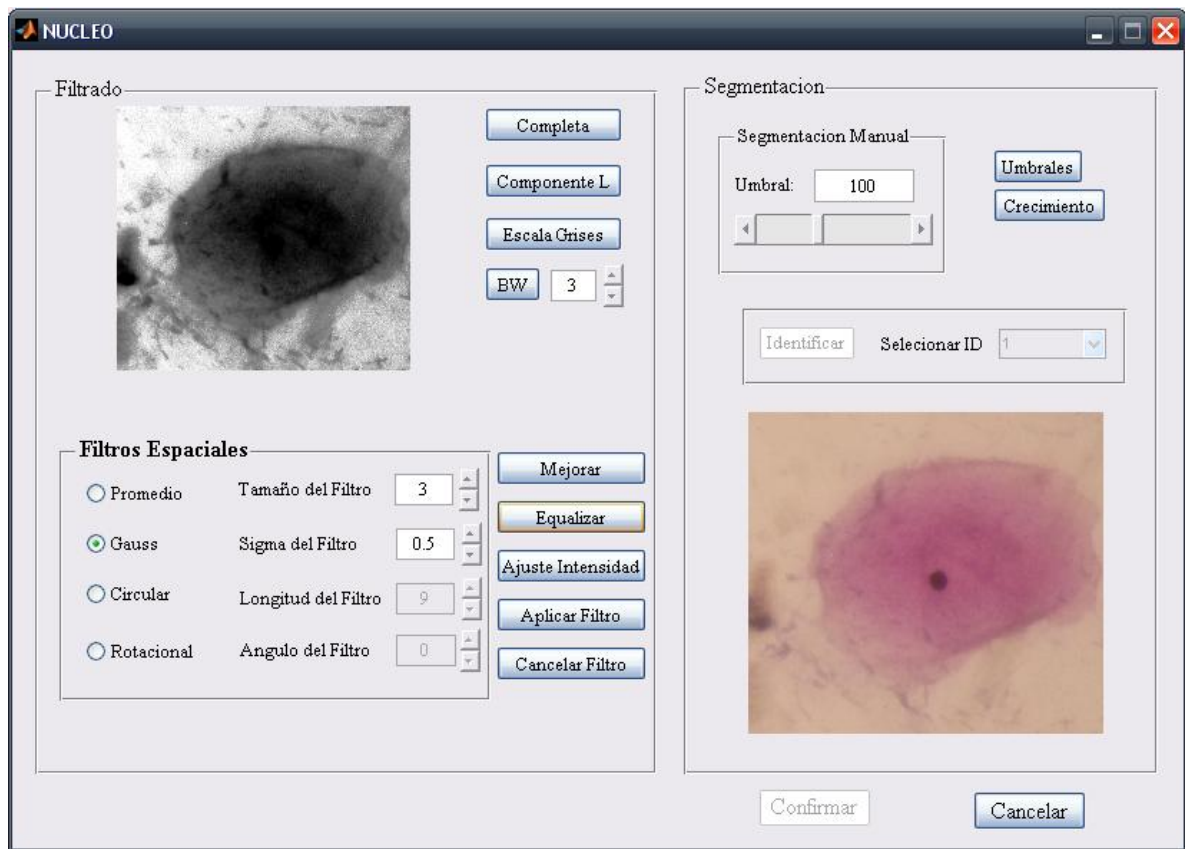


Figura 42. Opciones aumento del contraste 2.

A continuación se muestra el resultado obtenido por el botón Ajuste Intensidad, es un tanto parecido a el efecto de Equalizar por el tipo de imagen, este varía según el tipo de imagen.

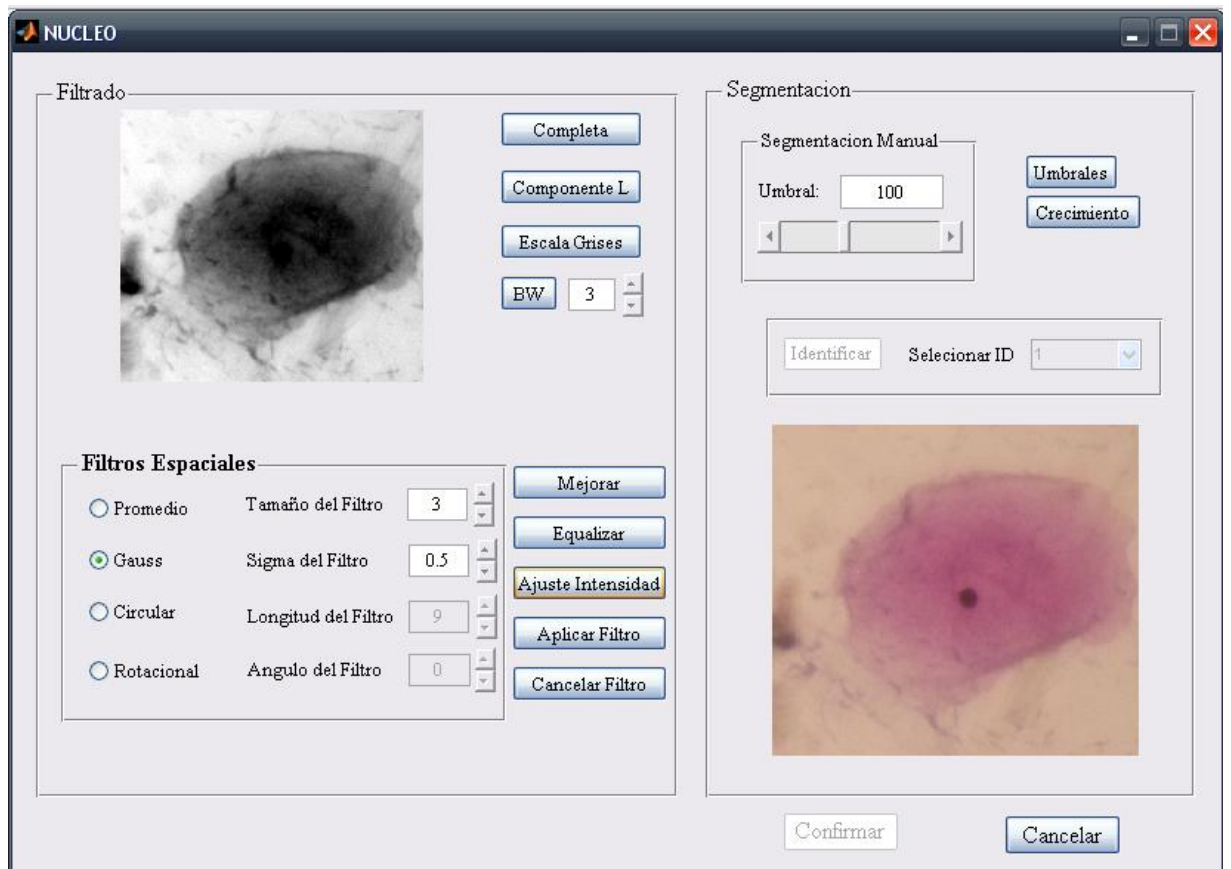


Figura 43. Opciones aumento del contraste 3.

Cuando se tiene una buena imagen con respecto al objeto que se quiere se procede a segmentar. Para el caso anterior funcionó muy bien el botón Mejorar por eso se escoge para segmentar el núcleo. La figura 44 muestra la segmentación del núcleo

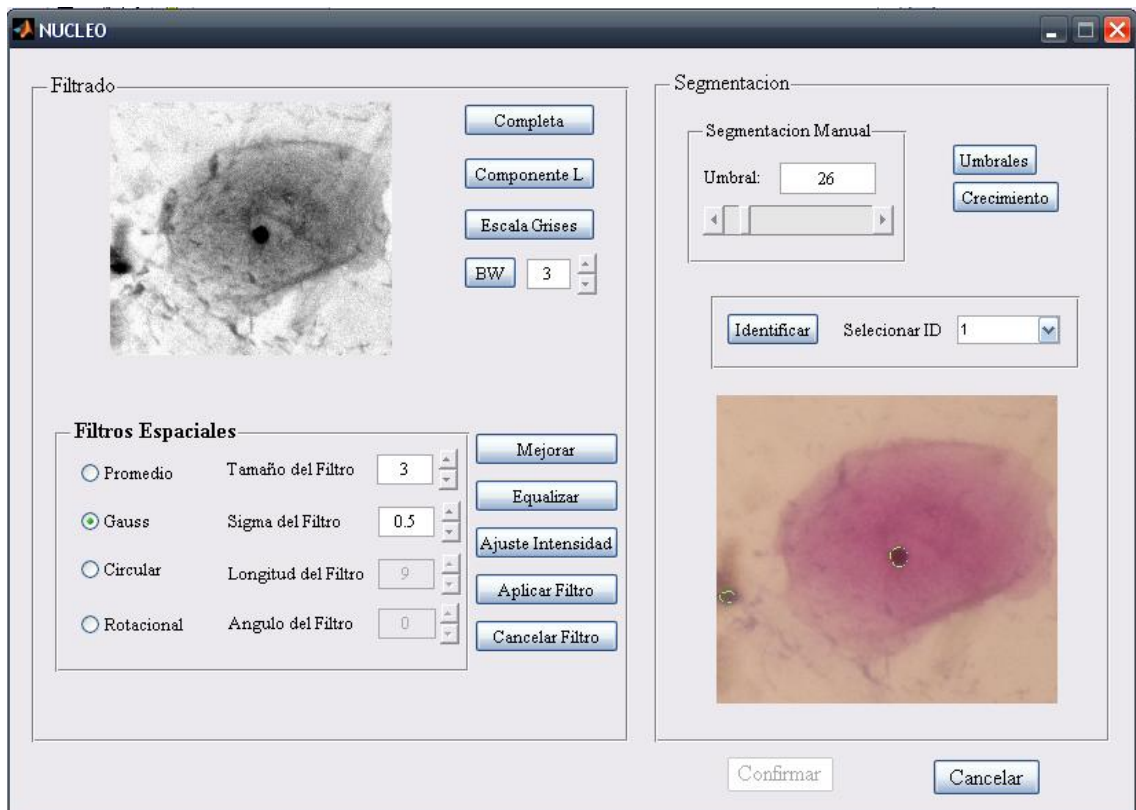


Figura 44. Segmentación del núcleo

En este paso se escoge segmentar por umbrales, crecimiento o manual. Si se escoge manual el punto de referencia es la barra deslizante la cual va a deslizarse hacia la derecha para aumentar el umbral. Una vez la imagen queda segmentada se procede a Identificar el área que se desea. La figura 45 y 46 muestran este paso.

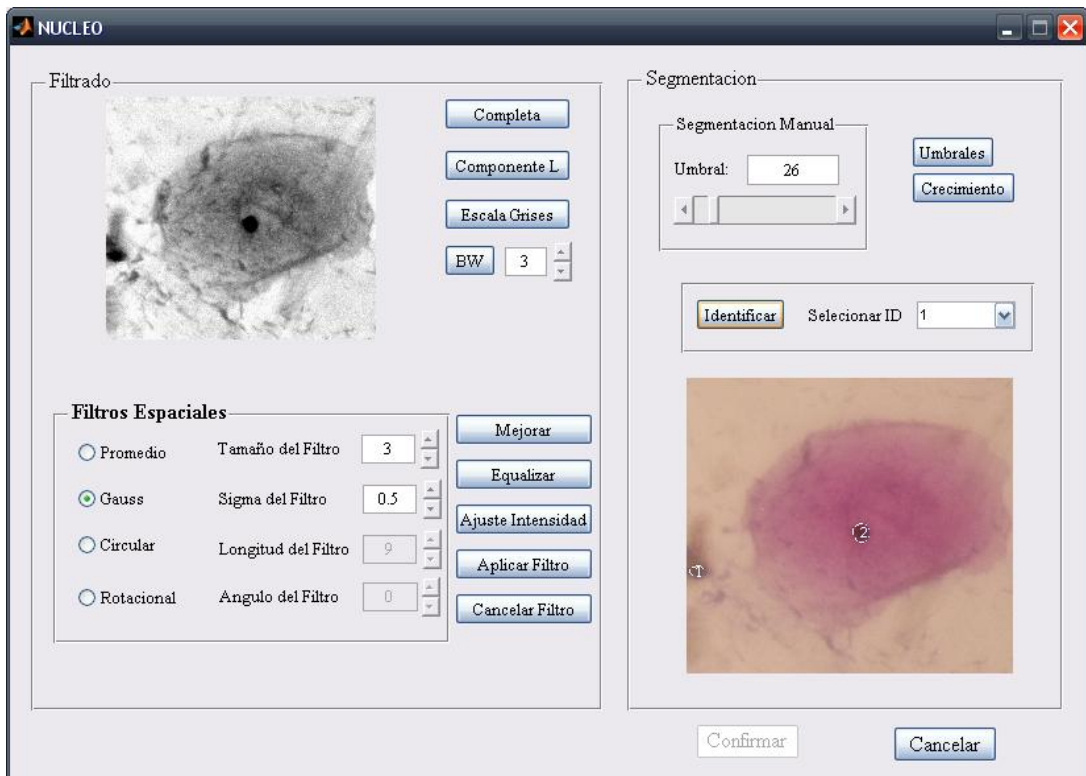


Figura 45. Identificar núcleo

Se selecciona el núcleo que se desea.

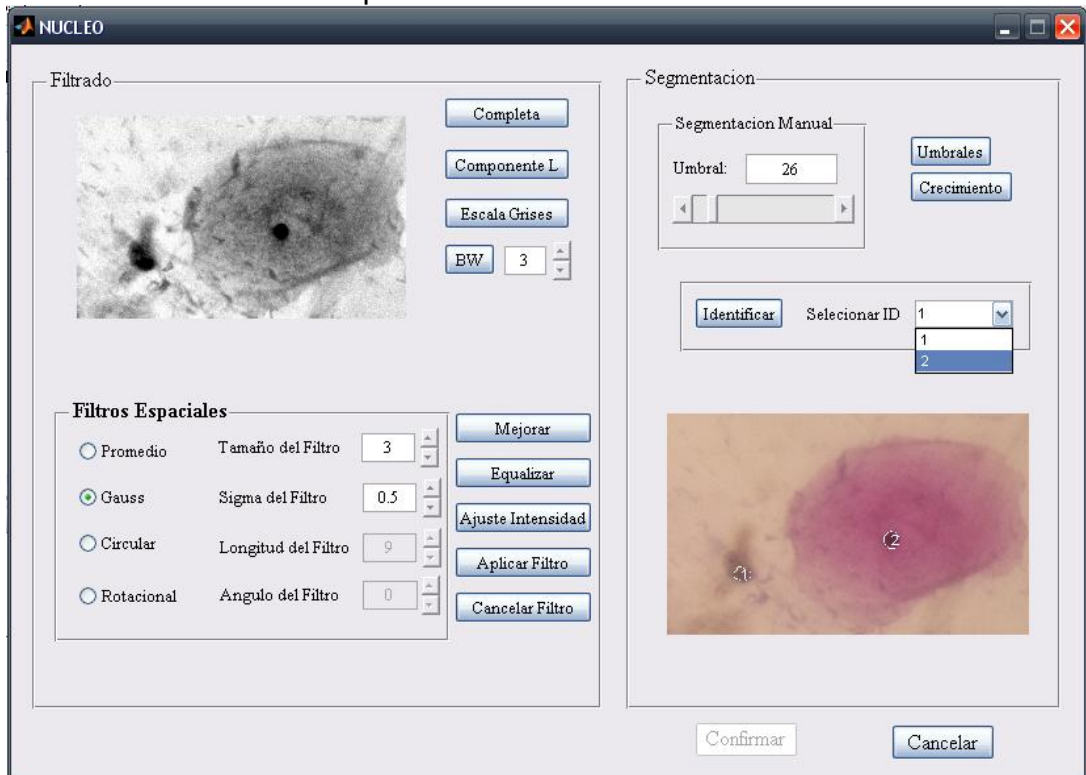


Figura 46. Seleccionar núcleo

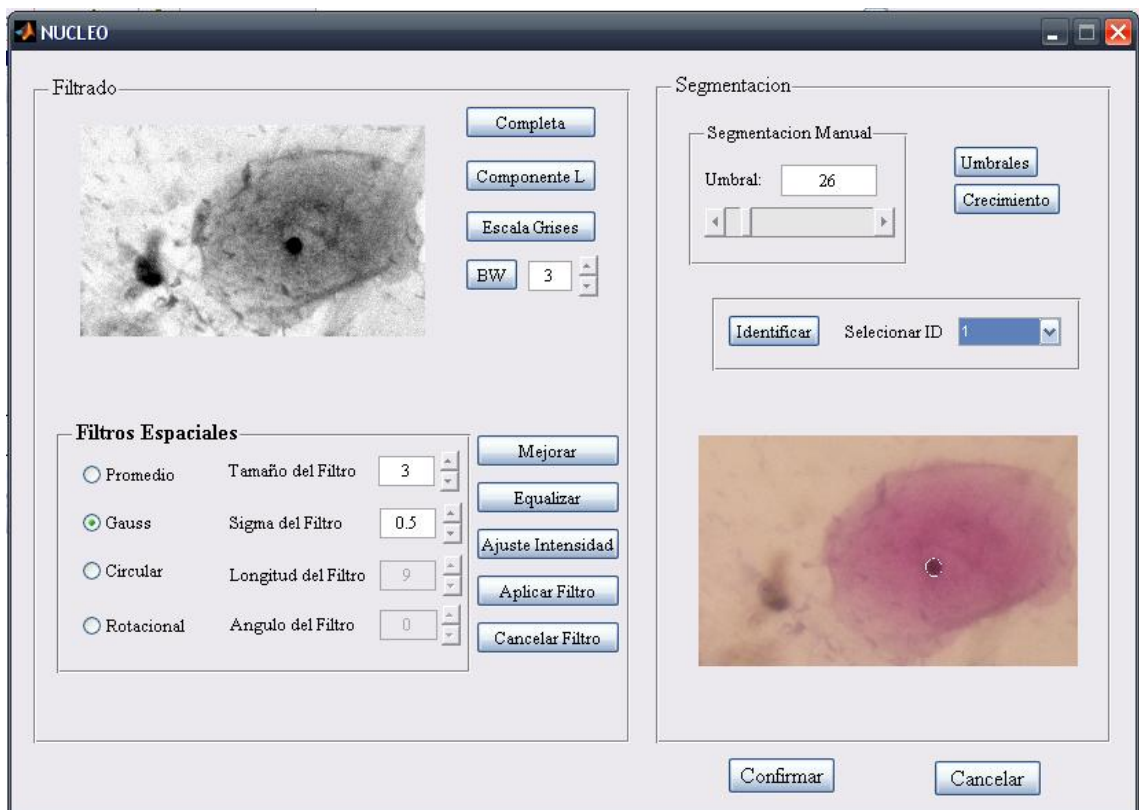


Figura 47. Núcleo segmentado

Una vez segmentado el núcleo, se confirma la selección.



Figura 48. Datos de las características del núcleo

Ahora se procede a al análisis del citoplasma; este es exactamente igual, con la única diferencia que tiene una opción más que se encarga de eliminar el ruido, esto con el objetivo de que el citoplasma quede bien segmentado. La siguiente figura muestra el proceso



Figura 49. Eliminación del ruido para el citoplasma

No es conveniente utilizar el botón Retoque, debido a que altera la imagen sobre todo para le etapa de segmentación. Es conveniente sólo en caso extremo con imágenes con bastante ruido.

Ahora se procede a segmentar de la misma forma que el núcleo obteniendo lo siguiente.

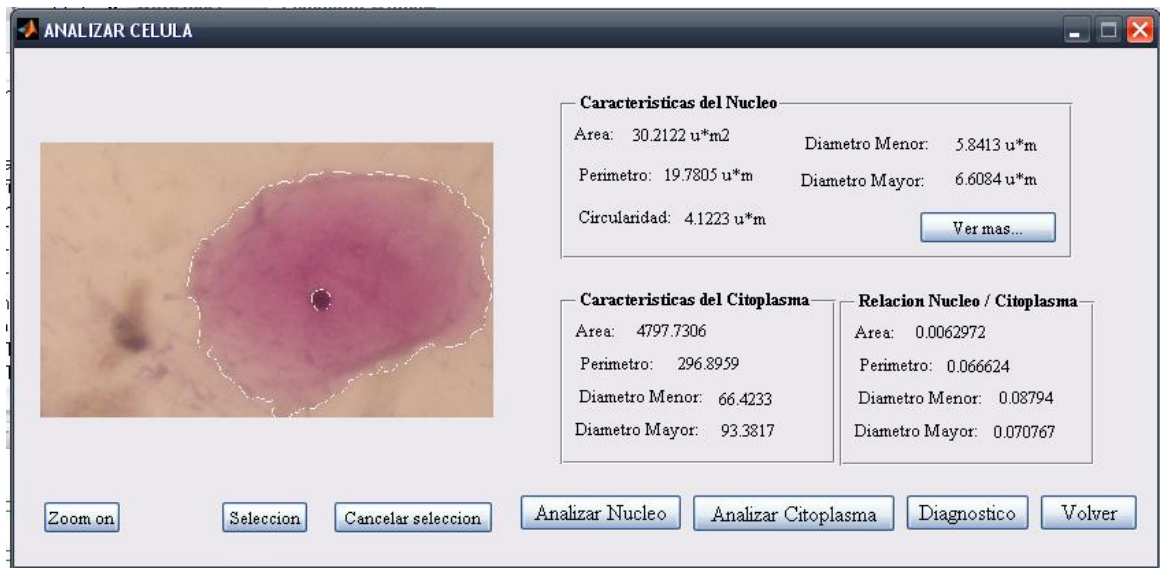


Figura 50. Núcleo y Citoplasma segmentados

Cuando se tiene tanto núcleo como citoplasma reconocidos totalmente se procede al diagnostico y la herramienta de clasificación dirá si la célula es normal o anormal.

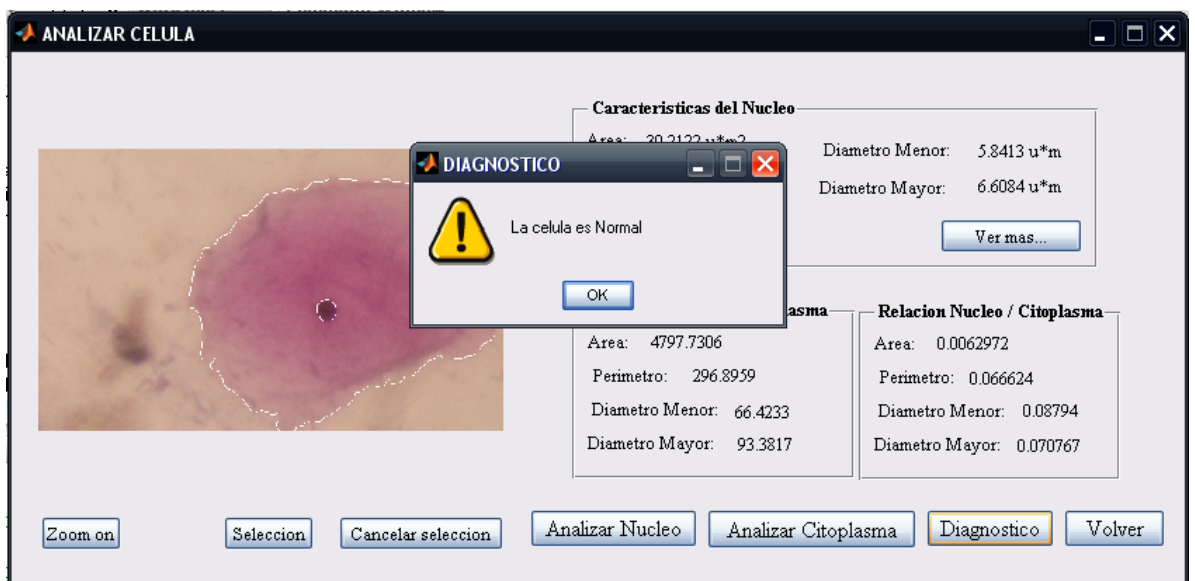


Figura 51. Diagnostico de la herramienta clasificadora.

En este caso el resultado fue célula normal, pero también está el otro caso, se muestra en la figura 52.



Figura 52. Diagnostico de la herramienta clasificadora 1

Al utilizar el botón CONFIGURAR.

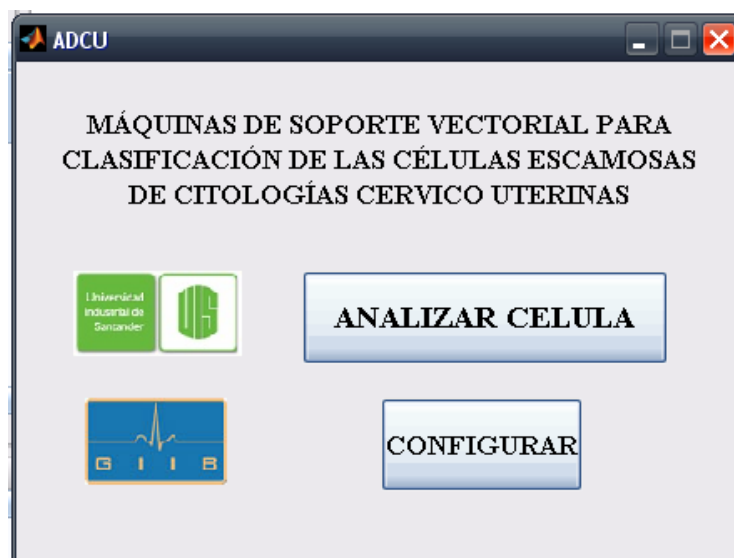


Figura 53. Ventana principal 2

Se abre la siguiente ventana

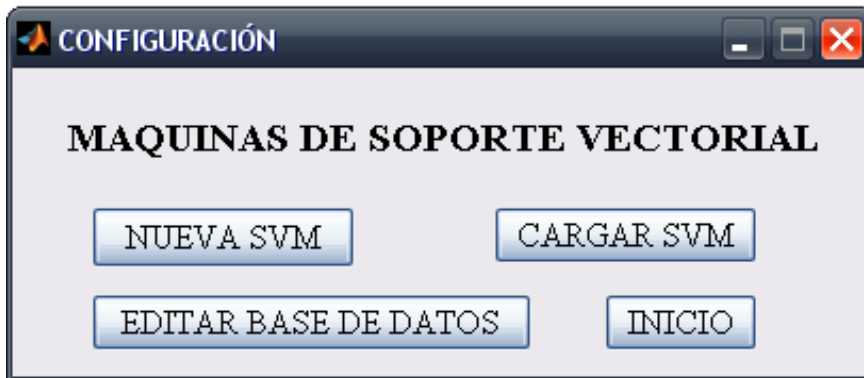


Figura 54. Ventana configuración

Con el botón NUEVA SVM se crea una nueva máquina de soporte vectorial, modificando las características. La figura 55 lo muestra.

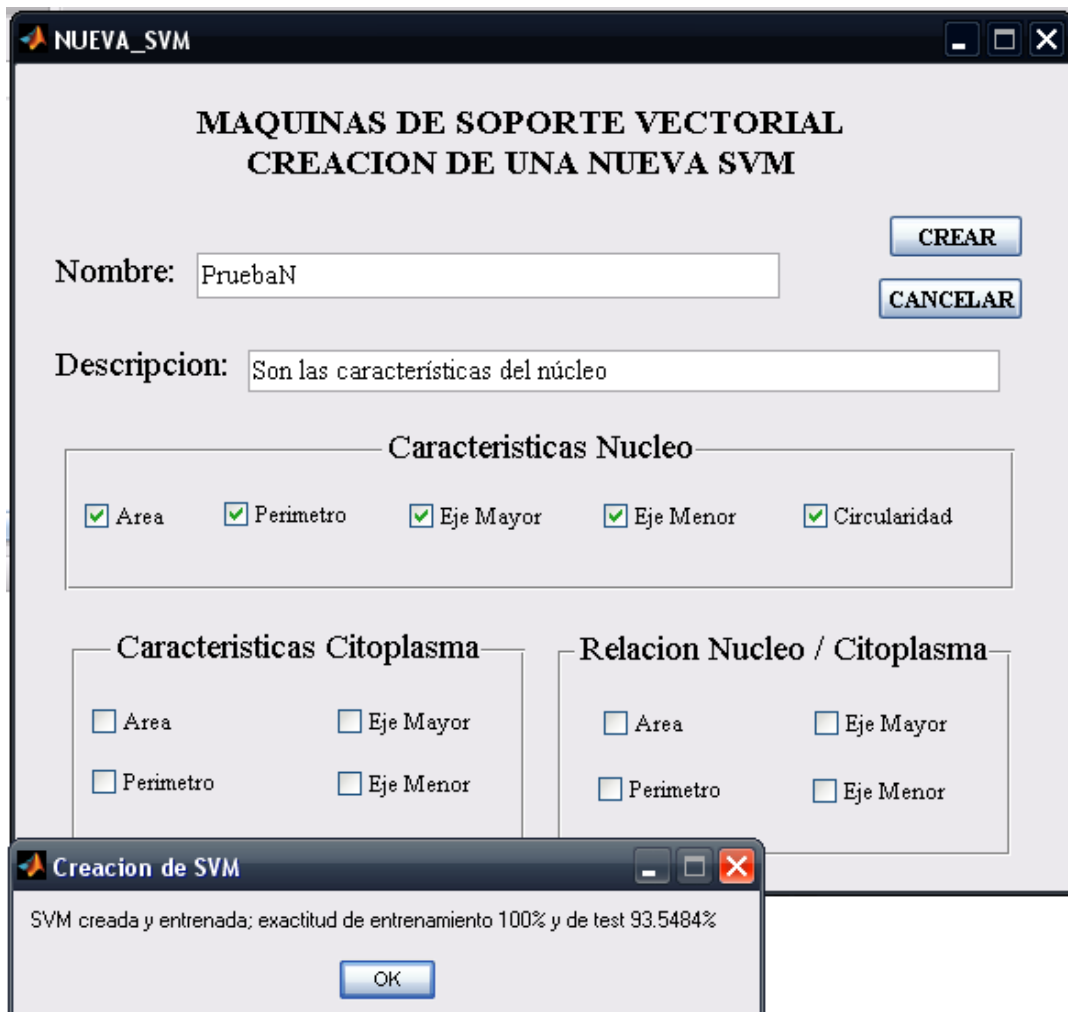


Figura 55. Creando una nueva SVM

En esta ventana es posible escoger el tipo de características que se desee. En el cuadro de texto de Nombre se le da el nombre a la nueva SVM; en el cuadro de texto Descripción describimos brevemente que características llevará la SVM a crear.

Luego de escoger las características deseadas, el nombre y descripción, procedemos a oprimir el botón CREAR, seguidamente la SVM comienza su entrenamiento. Al terminar de entrenar aparece una pequeña ventana tal como se muestra en la figura 55 donde se muestra que la SVM fue creada y entrenada, además muestra el porcentaje de exactitud de entrenamiento y prueba.

Una vez se acepta, se abre la siguiente ventana



Figura 56. SVM creada y puesta en el sistema

Como se puede observar en la figura 56, se muestra la SVM creada con el nombre PruebaN. Si se desea que la SVM creada quede en el sistema simplemente se oprime el botón CARGAR SVM tal como se muestra en la figura 57.



Figura 57. Cargar SVM

Se debe tener cuidado al escoger cargar la SVM, por ejemplo si esta posee una exactitud de entrenamiento y prueba bastante bajas no es conveniente cargarla.

Con el botón CAMBIAR SVM se abre una SVM creada y guardada previamente en un archivo. La siguiente figura muestra el procedimiento.

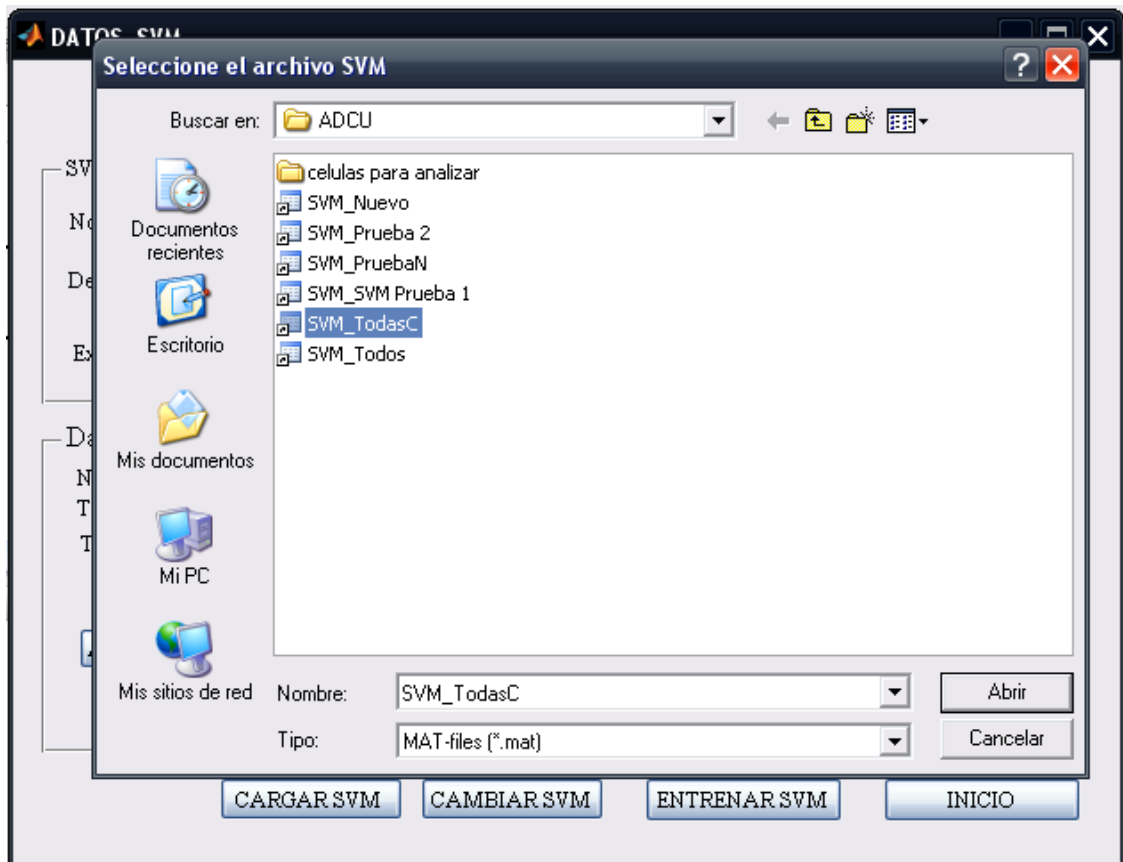


Figura 58. Cambiar SVM

En este ejemplo se abre la SVM llamada SVM_TodasC. Una vez se da abrir sale la siguiente ventana. En esta se muestra una pequeña ventana donde dice que el archivo fue cargado con éxito y como se puede observar la SVM cambiada entra en el sistema pero para que quede en el por defecto, nuevamente se oprime el botón CARGAR SVM.



Figura 59. Cambio de la SVM

Botón EDITAR BASE DE DATOS

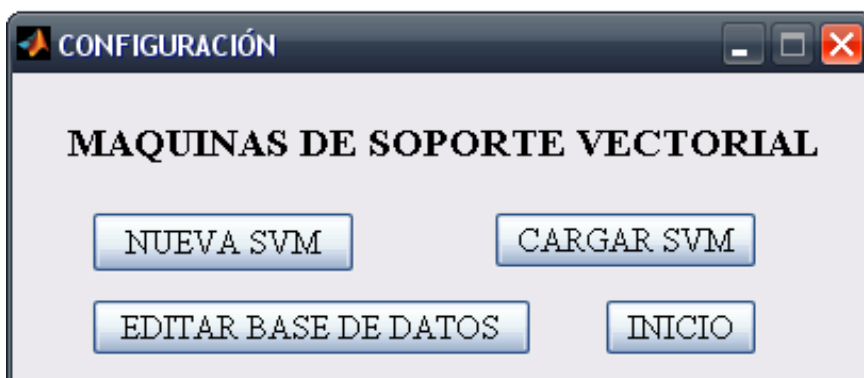


Figura 60. Ventana configuración 2

Al abrir la ventana de edición de la base de datos, es posible agregar o eliminar una nueva imagen digital ya se de células normales o anormales. Se debe tener en cuenta que al editar la base de datos se debe tener pleno conocimiento y seguridad que las células sean normales o anormales. El software por medio de una ventana preguntará si la célula es normal o anormal.

La figura 61 muestra la ventana editar base de datos.



Figura 61. Base de Datos

Como se había comentado anteriormente, es posible editar la base de datos con los botones AGREGAR y ELIMINAR.

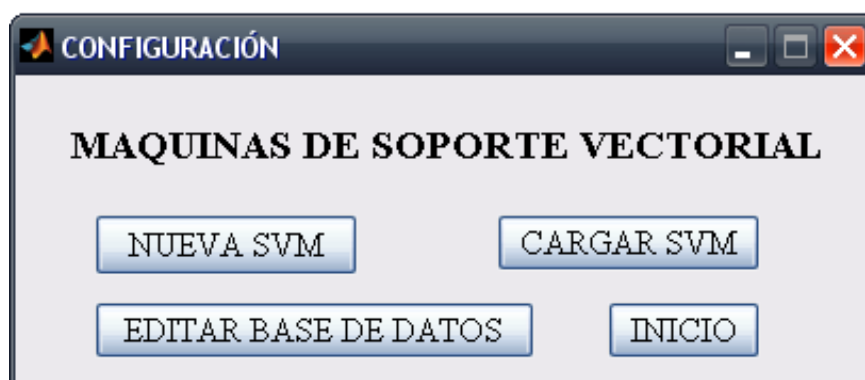


Figura 62. Ventana configuración 3

Al oprimir el botón cargar SVM, se puede cargar la SVM directamente si necesidad de entrar a editar la base de datos. La figura 63 muestra la carga de la SVM.

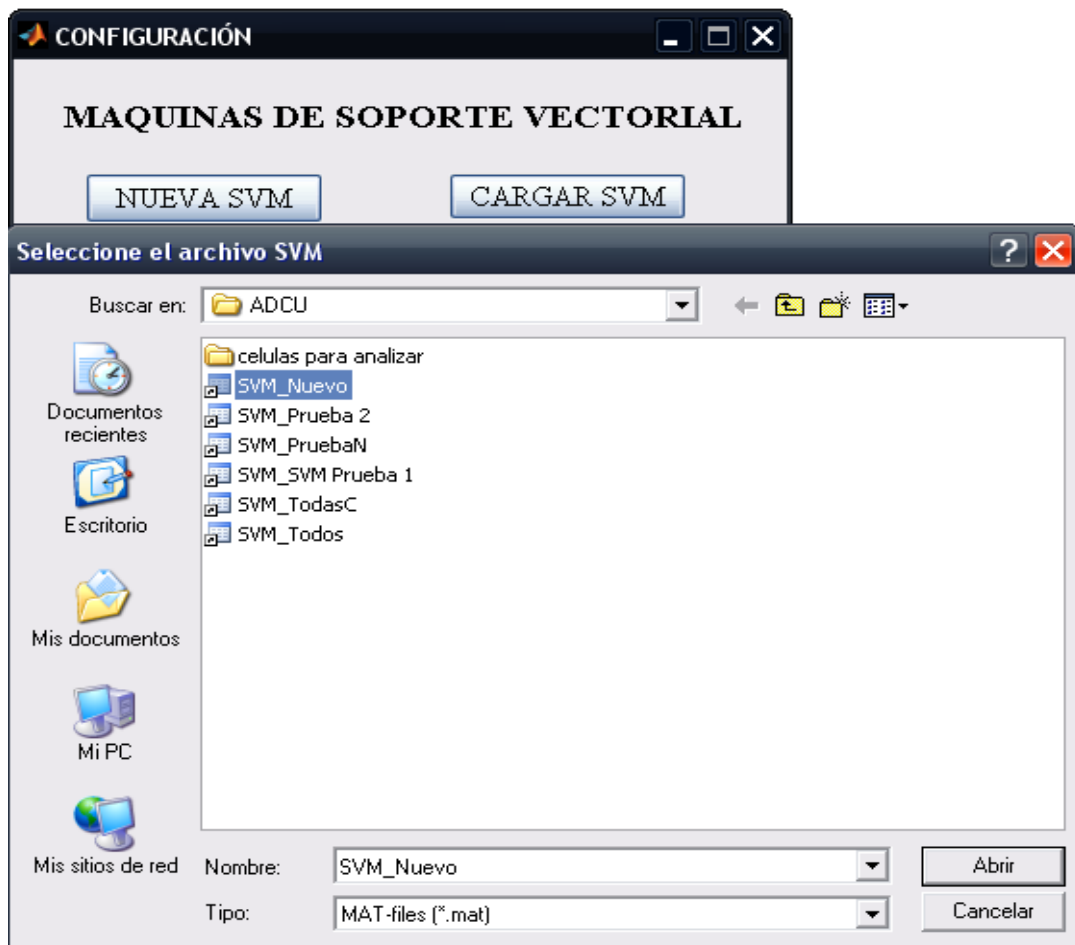


Figura 63. Carga de la SVM directamente

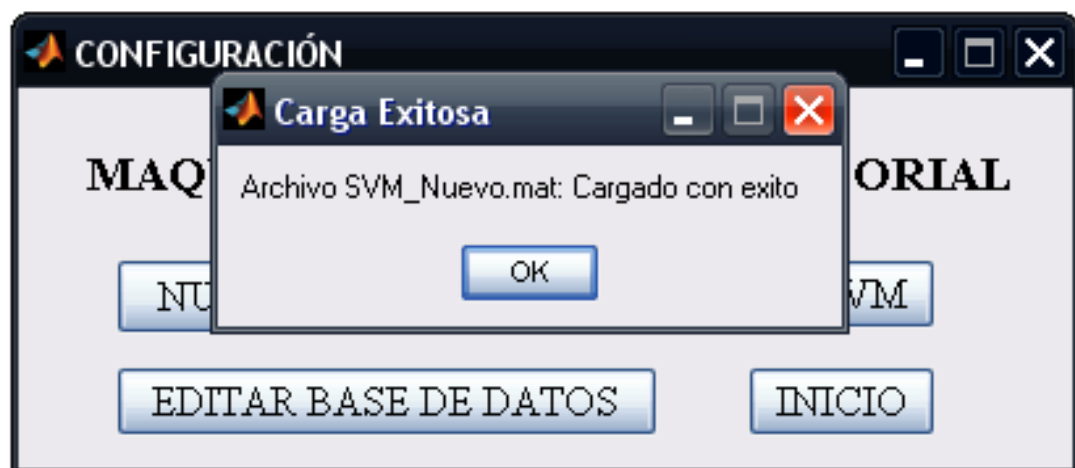


Figura 64. Archivo SVM cargado con éxito

El botón inicio carga nuevamente la ventana principal de la herramienta software.