

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA



**ESTUDIO COMPARATIVO DE TÉCNICAS DE INTELIGENCIA
ARTIFICIAL PARA EVALUAR VIABILIDAD EN CÉLULAS DE
CÁNCER DE MAMA CULTIVADAS IN VITRO.**

T E S I S

que presenta para obtener el grado de MAESTRO EN CIENCIAS E INGENIERIA

PABLO HERNANDEZ JARAMILLO

**DIRECTOR DE TESIS:
DR. DANIEL HERNANDEZ BALBUENA**

MEXICALI, B. C.

JUNIO DE 2022

TESIS MAESTRIA

ESTUDIO COMPARATIVO DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL PARA EVALUAR VIABILIDAD EN CÉLULAS DE CÁNCER DE MAMA CULTIVADAS IN VITRO.

AUTOR: Pablo Hernández Jaramillo
DIRECTOR: Daniel Hernández Balbuena
CO-DIRECTOR: Angélica López Izquierdo

Comité de tesis nombrado el día 30 de enero del 2022.

DIRECTOR Dr. Daniel Hernández Balbuena

CODIRECTOR Dra. Angélica López Izquierdo

MIEMBRO 1 Dra. Wendy Flores Fuentes

MIEMBRO 2 Dr. Oleg Sergiyenko

MIEMBRO 3 Dra. Blanca Dalila Pérez Pérez

ALUMNO Pablo Hernández Jaramillo

PRESIDENTE DEL SACC Dr. Julio Cesar Rodríguez Quiñonez

Coordinador de Investigación y Posgrado Dra. Wendy Flores Fuentes

Realizado el acto de defensa y lectura de tesis el día de junio del 2022

En el laboratorio de Electrónica de la Facultad de Ingeniería Mexicali.

Dedicatoria

A mi esposa: **Nélida Suhey Flores Ruvalcaba**

A mi hijo: **Oliver Issac Hernández Flores**

A mi abuela: **Blanca Celia Figueroa**

A mi abuelo: **Víctor Jaramillo Félix**

A mi madre: **Natalia Elvira Jaramillo Figueroa**

A mi maestro y amigo: **Dr. Daniel Hernández Balbuena**

Agradecimientos

Primero, quiero agradecer a YHWH, por brindarme el tiempo y el conocimiento durante esta etapa de mi vida y siempre suplir a mi familia.

Quiero agradecer a mi esposa Nélida Suhey Flores Ruvalcaba y a mi hijo Oliver Isaac Hernández Flores, por vivir esta etapa conmigo, por tolerarme y tener paciencia, por darme tiempo para continuar.

A mi mamá Natalia Elvira Jaramillo Figueroa, por animarme a continuar con mi preparación académica.

A mis abuelos Blanca Celia Figueroa y Víctor Jaramillo Félix, por apoyarme en mi preparación académica, por sus consejos y por brindarme la oportunidad de estudiar y enseñarme diversos oficios.

Al Dr. Daniel Hernández Balbuena por brindarme la oportunidad de realizar un posgrado y asesorarme durante mis estudios de licenciatura y posgrado.

Agradecimientos Institucionales

Agradezco a las instituciones que brindaron su apoyo durante el desarrollo de esta tesis:

Consejo Nacional de Ciencia y Tecnología (CONACyT)

Secretaría de Educación Pública (SEP)

Universidad Autónoma de Baja California (UABC)

RESUMEN de la Tesis de Pablo Hernández Jaramillo, presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS EN EL CAMPO DE CONOCIMIENTO DE ELÉCTRICA Mexicali, Baja California, México. Junio de 2022.

**ESTUDIO COMPARATIVO DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL
PARA EVALUAR VIABILIDAD EN CÉLULAS DE CÁNCER DE MAMA
CULTIVADAS IN VITRO**

Resumen aprobado por:

Dr. Daniel Hernández Balbuena
Director de tesis

En el presente trabajo de tesis se desarrolla un sistema para identificar y determinar la viabilidad de células de cáncer de mama, además de identificar restos celulares en fotografías tomadas del microscopio. El sistema utiliza técnicas de procesamiento digital de imágenes y algoritmos de inteligencia artificial para la identificación y evaluación de la viabilidad de las células. Se realiza una comparación entre técnicas de inteligencia artificial con el objeto de determinar la más eficiente. Para el desarrollo del sistema, se considera lo siguiente:

1. Un conjunto de datos de fotografías tomadas del microscopio, de cultivos de la línea celular de cáncer de mama MDA-MB-231.
2. Se aplican filtros espaciales con el propósito de conservar los contornos, eliminar ruido y homogeneizar la imagen del cultivo celular.
3. Se extraen características morfológicas de las células con el objetivo de generar vectores de información, para encontrar patrones.
4. Se utilizan tres modelos de aprendizaje supervisado, con el objetivo de discriminar las células viables y no viables, además, de identificar restos celulares, estos son bosque aleatorio, Bayes ingenuo e impulso adaptativo.
5. Se evalúan los algoritmos con el propósito de identificar el modelo de aprendizaje supervisado que presente un mejor desempeño.

ABSTRACT of the thesis, presented by (candidate name), in order to obtain the MASTER of ENGINEERING DEGREE in ELECTRONICS ENGINEERING. Mexicali, Baja California, México. June, 2022.

ESTUDIO COMPARATIVO DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL PARA
EVALUAR VIABILIDAD EN CÉLULAS DE CÁNCER DE MAMA CULTIVADAS IN
VITRO

Approved by:

Dr. Daniel Hernández Balbuena
Thesis Advisor

In this thesis work, a system is developed to identify and from the viability of breast cancer cells, in addition to identifying remains of cell photographs taken from a microscope. The system uses digital image processing techniques and artificial intelligence algorithms for the identification and evaluation of cell viability. A comparison is made between artificial intelligence techniques in order to determine the most efficient. For the development of the system, the following is considered:

1. A dataset of microscopic photographs of cultures of the breast cancer cell line MDA-MB-231.
2. Spatial filters are applied in order to preserve the contours, eliminate noise and homogenize the image of the cell culture.
3. Morphological characteristics of the cells are extracted in order to generate information vectors, to find patterns.
4. Three supervised learning models are used, with the aim of discriminating viable and non-viable cells, in addition to identifying cell remains, these are random forest, ingenious Bayes and adaptive impulse.
5. The algorithms are evaluated with the purpose of identifying the supervised learning model that presents the best performance.

Índice General

Dedicatoria	I
Agradecimientos	II
Agradecimientos institucionales	III
Resumen	IV
Abstract	V
Índice de Figuras	VI
Índice de Tablas	VII

ÍNDICE

1 Introducción	17
1.1. Viabilidad celular.....	17
1.2. Organización de la tesis	19
2. Algoritmos de inteligencia artificial y procesamiento de imágenes	20
2.1 Revisión de la literatura en la detección de células.	20
2.2 Fundamentos del procesamiento digital de imágenes.....	22
2.3 Etapas del procesamiento digital de imágenes.....	23
2.3.1 Filtros espaciales para imágenes	25
2.3.2 Filtro de difusión anisotrópica	26
2.3.3 Filtro Kuwahara	29
2.3.4 Filtro de media	31
2.4 Procesamiento digital de imágenes usando FIJI®	32
2.4.1 Menú procesar.....	34
2.4.2 Menú analizar.....	40
2.4.3 Menú complementos.....	46
2.5 Algoritmos de Inteligencia artificial para imágenes	51
2.5.1 Etapas de la aplicación de algoritmos de inteligencia artificial	53
2.6 Naive Bayes	54
2.6.1 Aprendizaje de Naive Bayes	55
2.7 Impulso adaptativo.....	57
2.7.1 Aprendizaje de impulso adaptativo.....	58
2.8 Bosque aleatorio.....	60
2.8.1 Aprendizaje de bosque aleatorio	61
2.9 Algoritmos de inteligencia artificial con WEKA®	62
3 Metodología.....	63
3.1 Metodología propuesta.....	63

3.2	Preprocesamiento de imágenes	65
3.3	Detección de objetos	66
3.4	Extracción de características.....	67
3.5	Entrenamiento	70
3.6	Clasificación	72
3.7	Evaluación de los clasificadores	74
4	Resultados.....	76
4.1	Conjunto de datos generado.....	76
4.2	Resultados obtenidos por procesamiento digital de imágenes.....	77
4.3	Resultados obtenidos al aplicar la extracción de características.....	82
4.4	Resultados obtenidos al aplicar los clasificadores	84
4.5	Comparativo de los resultados de aplicar los clasificadores.....	92
5.	Conclusiones.....	94
5.1	Líneas futuras de investigación.....	95
	APÉNDICE.....	96
	Referencias bibliográficas.....	102

Índice de Figuras

<u>Figura</u>		<u>Página</u>
<u>1</u>	Imágenes de la línea celular MDA-MB-231. a) Célula viable. (b) Célula no viable.	<u>18</u>
<u>2</u>	Representación gráfica de una imagen digital de células de cáncer de mama de 1980 x 1904 píxeles, el indicador de la intensidad en ese punto es $f(x, y)$.	<u>23</u>
<u>3</u>	Esquema general de las etapas del procesamiento digital de imágenes.	<u>25</u>
<u>4</u>	Efecto de filtrado de difusión anisotrópica en imágenes de células de cáncer de mama. a) Imagen original, b) imagen con 20 iteraciones del filtro (centro) y c) imagen con 60 iteraciones.	<u>28</u>
<u>5</u>	Efecto del uso del filtro Kuwahara (a) Imagen original, (b) imagen con máscara 7x7, (c) imagen con máscara 15x15.	<u>29</u>
<u>6</u>	Máscara de 3x3 con filtro promedio.	<u>31</u>
<u>7</u>	Efecto del uso del filtro de media (a) Imagen original, (b) imagen con máscara 2x2, (c) imagen con máscara 5x5.	<u>31</u>
<u>8</u>	GUI de FIJI®, y su barra de menú.	<u>32</u>
<u>9</u>	Logotipo de software FIJI®.	<u>33</u>
<u>10</u>	Menú procesar del software FIJI®.	<u>34</u>
<u>11</u>	Efecto de la aplicación del filtro de suavizado en una imagen de células de cáncer de mama. (a) Imagen original, (b) imagen con filtro de suavizado (Smooth).	<u>35</u>
<u>12</u>	Máscara 3x3 de filtro de agudizado.	<u>35</u>
<u>13</u>	Efecto de filtro de Agudizado (Sharpen) en una imagen de células de cáncer de mama. Imagen original (a), (b) imagen con filtro de Agudizado (Sharpen).	<u>36</u>
<u>14</u>	Máscara 3x3 de barrido horizontal (izquierda), máscara 3x3 de barrido vertical (derecha).	<u>36</u>
<u>15</u>	Efecto de aplicar el detector de bordes (Find Edges) en una imagen de cáncer de mama Imagen original (a), imagen con detector de bordes (Find edges) del menú procesar (b).	<u>37</u>

<u>16</u>	Efecto de aplicar el detector de máximos locales (Find maxima) en una imagen de cáncer de mama Imagen original (a), imagen con máximos locales (Find maxima) (b), Figura umbralizado con los máximos locales (c).	<u>37</u>
<u>17</u>	Tabulación de parámetros de análisis de imágenes.	<u>38</u>
<u>18</u>	Efecto de aplicar la corrección de contraste (enhance contrast) en una imagen de células cáncer de mama. (a) Imagen original, (b) imagen con corrección de contraste (Enhance contrast).	<u>39</u>
<u>19</u>	Menú analizar del software FIJI®.	<u>40</u>
<u>20</u>	Efecto de aplicar el análisis de partículas (Enhance Contrast) en una imagen de células de cáncer de mama en base a los parámetros tamaño y la circularidad. (a) Imagen original, (b) imagen analizada con tamaño 50-1000 y circularidad 0-0.60, (c) imagen analizada con tamaño 50-1000 y circularidad 0-0.10.	<u>42</u>
<u>21</u>	Submenú establece medidas del software FIJI®.	<u>43</u>
<u>22</u>	Menú complementos del software FIJI®.	<u>46</u>
<u>23</u>	Proceso de Segmentación Entrenable de Weka (TWS).	<u>47</u>
<u>24</u>	Interfaz gráfica de usuario de Segmentación Entrenable de Weka (TWS).	<u>50</u>
<u>25</u>	<u>imagen</u> original (a), imágenes resultantes de Segmentación Entrenable de Weka (TWS), la imagen segmentada (b) y el mapa de probabilidades de la imagen (c).	<u>50</u>
<u>26</u>	Ventana de ajustes de conFiguración de TWS	<u>51</u>
<u>27</u>	Proceso general en la aplicación de algoritmos de inteligencia artificial.	<u>53</u>
<u>28</u>	Resultados de aplicar el algoritmo de Naive Bayes al conjunto de datos “Iris”	<u>55</u>
<u>29</u>	Secuencia del modelo, primer algoritmo “débil” (izquierda), segundo algoritmo “débil” (derecha).	<u>57</u>
<u>30</u>	Secuencia del modelo de bosque aleatorio, con cada árbol de decisión.	<u>60</u>
<u>31</u>	Proceso de aprendizaje del algoritmo de bosque aleatorio.	<u>61</u>
<u>32</u>	Interfaz gráfica de usuario (GUI) del software WEKA®.	<u>63</u>
<u>33</u>	Metodología propuesta, la cual consta de cinco etapas, el preprocesamiento de datos, la detección de objetos, la extracción de características, el entrenamiento y la clasificación.	<u>64</u>

<u>34</u>	Metodología implementada en el preprocesamiento digital de imágenes celulares, para la línea celular MDA-MB-231.	<u>65</u>
<u>35</u>	Metodología implementada en la detección de bordes.	<u>66</u>
<u>36</u>	Metodología empleada en la realización del enmascarado.	<u>66</u>
<u>37</u>	Características morfológicas extraídas mediante el software FIJI®.	<u>69</u>
<u>38</u>	Metodología empleada en la extracción de características morfológicas de las células.	<u>70</u>
<u>39</u>	Metodología empleada en el entrenamiento del algoritmo en TWS.	<u>71</u>
<u>40</u>	Imágenes etiquetadas y clasificadas mediante mapa de probabilidades y segmentación	<u>73</u>
<u>41</u>	Imágenes fragmentadas e imágenes clasificadas con cada clase designada.	<u>73</u>
<u>42</u>	Ejemplo de una matriz de confusión de 6x6.	<u>74</u>
<u>43</u>	Conjunto de imágenes originales de la base de datos.	<u>77</u>
<u>44</u>	Imágenes con filtro de difusión anisotrópica.	<u>78</u>
<u>45</u>	Imágenes con filtro de agudizado.	<u>79</u>
<u>46</u>	Conjunto de imágenes con filtro Kuwahara.	<u>80</u>
<u>47</u>	Imágenes con filtro Sobel.	<u>81</u>
<u>48</u>	Conjunto de imágenes con extracción de características.	<u>83</u>
<u>49</u>	Conjunto de imágenes clasificadas, resultantes del algoritmo impulso adaptativo.	<u>86</u>
<u>50</u>	Conjunto de imágenes clasificadas, resultantes del algoritmo Bayes ingenuo.	<u>89</u>
<u>51</u>	Conjunto de imágenes clasificadas, resultantes del algoritmo bosque aleatorio.	<u>91</u>
<u>52</u>	Gráfica comparativa de los tres algoritmos de la metodología propuesta.	<u>92</u>
<u>53</u>	Método para aplicar filtros espaciales en FIJI®.	<u>96</u>
<u>54</u>	Método para aplicar el operador gradiente de Sobel en TWS.	<u>96</u>
<u>55</u>	Método para extraer las características morfológicas en FIJI®	<u>97</u>
<u>56</u>	Método para inicializar TWS y generar etiquetas.	<u>98</u>
<u>57</u>	Método para asignar clases y conFigurar el algoritmo de I.A.	<u>98</u>
<u>58</u>	Método para entrenar algoritmo en TWS, e imagen resultante del entrenamiento.	<u>99</u>
<u>59</u>	Método para clasificar múltiples imágenes en TWS.	<u>100</u>

<u>60</u>	Método para visualizar los vectores de datos en WEKA®	<u>100</u>
<u>61</u>	Método para obtener la matriz de confusión y sus métricas en WEKA®.	<u>101</u>

Índice de Tablas

<u>Tabla</u>		<u>Página</u>
<u>1</u>	Parámetros de filtro de difusión anisotrópica.	<u>28</u>
<u>2</u>	Modelos de aprendizaje supervisado (izquierda) y modelos de aprendizaje no supervisados (derecha).	<u>52</u>
<u>3</u>	Características morfológicas de las células de cáncer de mama y su descripción	<u>67</u>
<u>4</u>	Tabla comparativa de los tres modelos de aprendizaje supervisado.	<u>82</u>
<u>5</u>	Matriz de confusión del algoritmo impulso adaptativo.	<u>84</u>
<u>6</u>	Métricas de desempeño del algoritmo de impulso adaptativo.	<u>85</u>
<u>7</u>	Matriz de confusión del algoritmo de Bayes ingenuo.	<u>88</u>
<u>8</u>	Métricas de desempeño del algoritmo de Bayes ingenuo.	<u>88</u>
<u>9</u>	Matriz de confusión del algoritmo de bosque aleatorio.	<u>90</u>
<u>10</u>	Métricas de desempeño del algoritmo de bosque aleatorio.	<u>90</u>
<u>11</u>	Tabla comparativa de los tres modelos de aprendizaje supervisado.	<u>92</u>

Capítulo 1

1 Introducción

1.1. Viabilidad celular

Los ensayos de citotoxicidad *in vitro* forman parte de la primera etapa en la evaluación de fármacos. Con base en los resultados de la evaluación citotóxica es posible obtener información de la biocompatibilidad del fármaco [1]. Con el objetivo de mantener una proliferación celular adecuada en las muestras usadas en la evaluación de fármacos, es necesario conocer la viabilidad celular del cultivo o de la línea celular con que se trabaje [2]. La viabilidad celular es definida como la cantidad de células vivas dentro de una muestra, y esta proporciona un control de calidad, con el que se encuentra el mejor momento para hacer uso del cultivo. Existen diversos métodos para evaluar la viabilidad celular, con los que se focaliza la proliferación de las células en la caja de Petri, los métodos utilizan como referencia las funciones celulares [3]. Estos métodos son mostrados a continuación:

- Evaluación de la integridad de la membrana celular.
- Ensayo por reducción del compuesto MTT (3 - 2,5-bromuro de Diphenyltetrazolium).
- Ensayos de cuantificación de los niveles de ATP (Adenosín Trifosfato).
- Ensayos de síntesis de ADN (Ácido Desoxirribonucleico).
- Métodos según la morfología celular.
- Métodos de la expresión génica,
- Método por energía dispersiva de rayos “X”

La aplicación de los métodos y ensayos para evaluar la viabilidad celular, generan gastos y consumen tiempo, que el experto podría invertir en analizar otras muestras. La subjetividad es una consecuencia del proceso ante el cual el experto se enfrenta, por lo que el mantener un control de calidad que le permita monitorear la viabilidad de los cultivos, a bajo costo y con un menor consumo de tiempo, sería de gran apoyo para la evaluación de la viabilidad de células en ambiente de laboratorio.

En la Figura 1, se muestran las fotografías tomadas al microscopio de dos células de la línea celular MDA-MB-231, una viable y otra no viable. En ellas se puede apreciar una leve diferencia en la morfología de la célula, e incluso en la membrana celular y en la distribución del citoplasma, sin embargo, no es visible si se encuentran adheridas al fondo de la caja de Petri, o si el núcleo está expuesto, además, la cantidad de células que se encuentran en el cultivo, hace aún más complicado distinguirlas. Por lo que es difícil discriminar las células viables y las que no lo son.

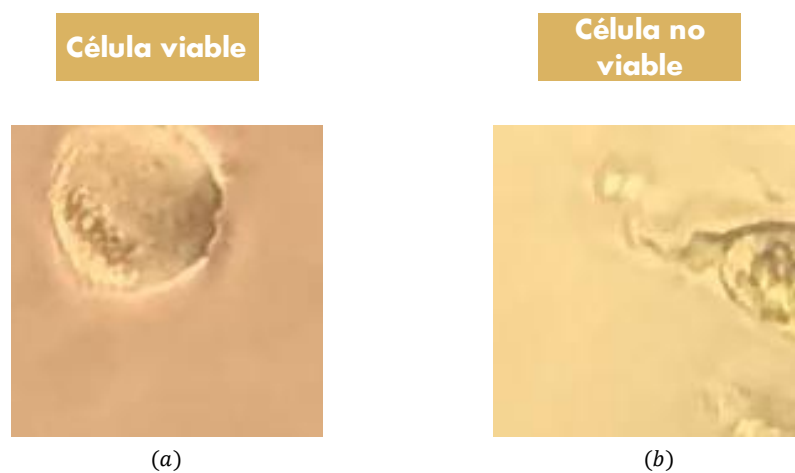


Figura 1: Imágenes de la línea celular MDA-MB-231. a) Célula viable. (b) Célula no viable.

Durante los últimos años, se han desarrollado una gran variedad de modelos de aprendizaje y técnicas de procesamiento digital de imágenes, que permiten detectar patrones visuales en fotografías, de forma eficiente y con bajo consumo de tiempo. Esto los hace elegibles para ser usados en la determinación de la viabilidad de células de cáncer de mama en ambientes de laboratorio.

El presente trabajo de tesis desarrolla una metodología que le permita a un experto analizar las muestras de una línea celular mediante visión computacional. En su desarrollo de este trabajo se usa la línea celular MDA-MB-231. Se comparan el desempeño de tres algoritmos de inteligencia artificial, para establecer el modelo de aprendizaje que arroje los resultados más óptimos. Además, este proyecto de tesis proporciona dos grados de apoyo al experto, estos son:

- Mejorar las condiciones visuales del cultivo, con el propósito de facilitar el ensayo o método de proliferación celular.
- Evaluar la viabilidad celular del cultivo, mediante modelos inteligentes, con el objetivo de proporcionar al experto, una segunda opinión o establecer el resultado mediante el algoritmo.

Por otro lado, la metodología utiliza dos softwares de licencia GNU (Licencia Pública General). Uno de procesamiento digital de imágenes, FIJI® y otro de inteligencia artificial, WEKA® (Waikato Environment for Knowledge Analysis).

1.2. Organización de la tesis

El presente trabajo de tesis: se divide se la siguiente manera:

- En el capítulo 1, se presenta una introducción al problema de determinar la viabilidad celular en ambientes de laboratorio, y la aplicación de la visión técnica para su determinación en imágenes histopatológicas.

- En el capítulo 2, se presenta una revisión de la literatura especializada, ahondando en los fundamentos de la teoría de procesamiento digital de imágenes y en los modelos de inteligencia artificial elegibles para determinar la viabilidad celular en imágenes de células de cáncer de mama.
- El capítulo 3, presenta la metodología propuesta de acuerdo con lo mencionado en el capítulo 2, además, describe cada una de las etapas, planteando las bases para exponer los resultados.
- En el capítulo 4, se presentan los resultados obtenidos al aplicar las técnicas de preprocesamiento digital de imágenes, la segmentación, la extracción de características y los algoritmos de clasificación, que son base de la metodología propuesta.
- Finalmente, el capítulo 5, presenta las conclusiones del presente trabajo de tesis y se exponen las líneas futuras de investigación.

Capítulo 2

Algoritmos de inteligencia artificial y procesamiento de imágenes

2.1 Revisión de la literatura en la detección de células.

Diversos trabajos de investigación se han llevado a cabo con el objetivo de detectar células de cáncer de mama mediante visión artificial, la mayoría de estos trabajos se enfocan en la modificación o aportación de las técnicas de procesamiento digital de imágenes o inteligencia artificial (en el caso de visión artificial), o en la metodología empleada (en el caso de que la técnica utilizada, sea análoga a la visión artificial). Sin embargo, el presente trabajo de tesis presenta una metodología similar a la de investigaciones previas, con una aplicación distinta, debido a que se busca evaluar la viabilidad celular en un cultivo, y no detectar células de cáncer de mama como tal. De acuerdo a diversos autores, las metodologías empleadas constan de las etapas que se mencionan a continuación:

- Procesamiento digital de imágenes.
- Fragmentación de áreas de interés en la imagen.
- Clasificación.

A continuación, se presentan las contribuciones del estado del arte de la presente investigación, algunas de las cuales se utilizaron como referencia.

En el artículo publicado por Luciana Gasparini en 2017, se realizó un estudio comparativo entre el software Cellprofiler® y el ensayo colorimétrico MTT, para evaluar la viabilidad celular en células de hepatoma (Hepa-1c1c7) y en fibroblastos (L929). En este trabajo el software Cellprofiler® utiliza un algoritmo que procesa la imagen de acuerdo a las especificaciones del investigador, además, le brinda información al experto sobre el tamaño y la morfología del cultivo celular, sin embargo, es necesario utilizar reactivos para diferenciar las células, aun así, utiliza el reactivo en una menor cantidad, en comparación a los métodos tradicionales, que suelen ser manuales [25].

En el caso del artículo publicado por Anuranjeeta en 2017, se realizó un estudio comparativo de softwares y complementos para el diagnóstico asistido por computadora, con el objetivo de detectar células cancerosas. La metodología empleada en este artículo fue la de utilizar un conjunto de datos existente, la cual fragmentaron y preprocesaron con filtros espaciales (para contraste), para segmentar y clasificar mediante los algoritmos Naive Bayes, Random Forest, Rotation Forest, J-Rip, MLP y LMT, las imágenes celulares. Realizando así la comparativa entre los algoritmos y los complementos de SIOX, RATS Y TWS [21].

En el año 2013, Christopher Sommer, describió las etapas generales para aplicar técnicas de procesamiento digital de imágenes, así como de aprendizaje automatizado, empleados en biología celular. Mencionando que la metodología general consta de cinco partes, las cuales son: Preprocesamiento de datos, detección de objetos, extracción de características, entrenamiento y finalmente la clasificación [5].

2.2 Fundamentos del procesamiento digital de imágenes

La imagen es una representación facsímil (idéntica o similar) de algún objeto, cosa o información no visible, como datos o señales. Con una perspectiva matemática, una imagen digital es una función en dos dimensiones con coordenadas espaciales con valor en magnitud para cada una de sus coordenadas, en forma discreta. Estas magnitudes son representadas con una función de la posición en la imagen $f(x, y)$ y su valor es proporcional a la intensidad de luz en ese punto de la escena captada en la imagen. Por lo que una imagen digital puede considerarse como una matriz bidimensional, en la que cada punto (x, y) es llamado pixel y contiene la información de la magnitud de la intensidad de luz, traducido en nivel de gris, o color (Figura 2).

Diversos dispositivos electrónicos pueden generar una imagen digital de una escena a partir de la división de la imagen en puntos (píxeles) y asignarles una magnitud (brillo o nivel de gris) y coordenada (x, y) . Esto es llevado a cabo por un dispositivo de carga acoplada o un dispositivo CMOS (Complementary Metal-Oxide-Semiconductor), tales dispositivos pueden ser cámaras digitales, escáneres o microscopios electrónicos. Las imágenes son discretizadas, ya que pasan por un convertidor analógico-digital y este proceso puede generar pérdida de información visual. La imagen digital puede variar según la resolución que contenga, la resolución es dada por la cantidad de píxeles por pulgada en la imagen [4]. Existen diversos formatos de imagen como JPG, PNG y TIFF. En su mayoría, los dispositivos que permiten obtener imágenes digitales se crea un registro de los colores, este puede ser generado a partir de una combinación cromática vectorial, como en el caso de la composición de colores en el formato RGB (Rojo, Verde, Azul) [4].

El procesamiento digital de imágenes puede definirse entonces como un conjunto de metodologías aplicadas a imágenes digitales con el propósito de adquirir, aislar, o mejorar información visual. Es así que al tomar una imagen se puede producir una versión modificada de la misma mediante operaciones matemáticas.

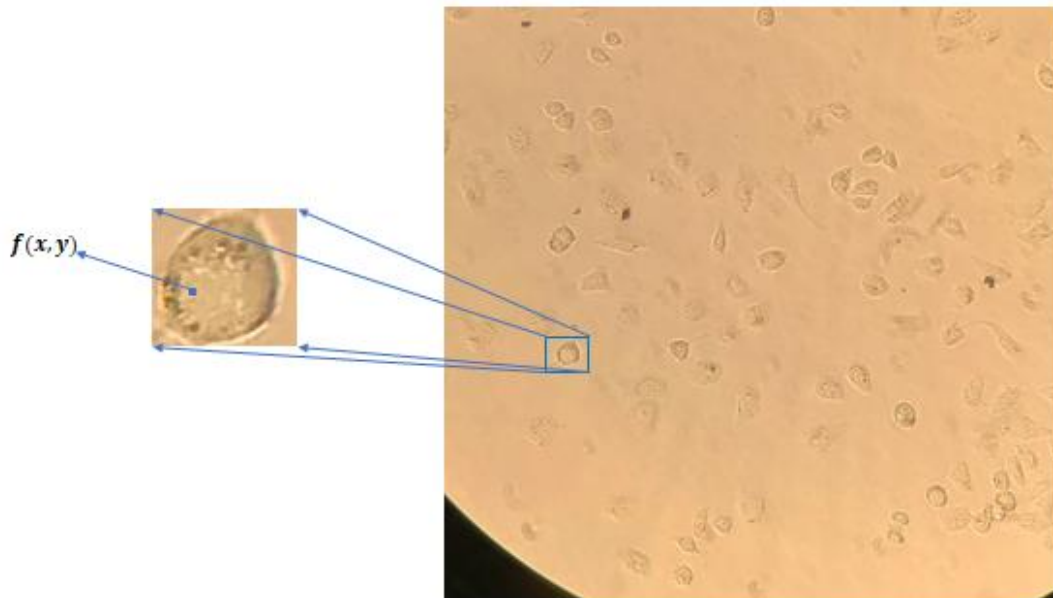


Figura 2: Representación gráfica de una imagen digital de células de cáncer de mama de 1980 x 1904 píxeles, el indicador de la intensidad en ese punto es $f(x, y)$.

2.3 Etapas del procesamiento digital de imágenes

El procesamiento digital de imágenes consta de diversas etapas, estas pueden variar según la problemática a tratar, sin embargo, existe un esquema general que comprende este conjunto de técnicas. El tratamiento de imágenes consta de cinco etapas, estas son la adquisición de la imagen, el preprocesamiento, la segmentación, representación y descripción, y en consecuencia la base de conocimiento de la imagen, las presentes etapas se muestran en la Figura 3 [5].

La adquisición de la imagen es la primera etapa del proceso del tratamiento de imágenes, para obtener una imagen digital es indispensable poseer algún dispositivo que permita digitalizar y discretizar la información visual o el objeto de interés, estos dispositivos pueden ser cámaras digitales o escáneres, esto por mencionar los más comunes, sin embargo, en el ámbito de la medicina existen dispositivos que pueden brindar una imagen digitalizada, como los rayos x, tomografías o ultrasonidos. Una vez que se ha obtenido la imagen digitalizada, se procede a preprocesar la imagen, esta es la segunda etapa del procesamiento digital de imágenes [4].

La segunda etapa es el preprocesamiento, su objetivo es el de obtener una imagen con una cantidad menor de impurezas y de ruido, en otras palabras, es el proceso que mejora la imagen para las siguientes etapas. Por lo general se utilizan técnicas de eliminación de ruido (premediación por filtrado) y de corrección del contraste (corrección de campo plano). La etapa posterior al preprocesamiento es la de la segmentación de la imagen.

La segmentación de una imagen es la tercera etapa del procesamiento de imágenes digitales, esta tiene el propósito de dividir la imagen en grupos de píxeles con características comunes, para obtener la forma y contorno de los objetos de interés. La segmentación es realizada a partir de distintas técnicas de umbralización (umbralizado de Otsu) y detección de bordes basados en gradiente (Prewitt, Sobel o Canny). Una vez que se ha segmentado la imagen, se procede a la cuarta etapa de este proceso, la cual es el reconocimiento e interpretación [4].

La cuarta etapa del procesamiento tiene el objetivo de etiquetar los objetos de interés y brindar una descripción y significado a los objetos que previamente se habían segmentado y etiquetado, esto es realizado por un usuario experto que continuamente trabaje con dichas imágenes.

La quinta y última etapa de este proceso es la base de conocimiento, y su propósito es el de recabar la información más relevante de las imágenes, obtenida a partir de las etapas previamente mencionadas, para limitar la búsqueda de información. Finalmente, el producto de la quinta puede representarse como una base de datos de imágenes de cierta índole [4].

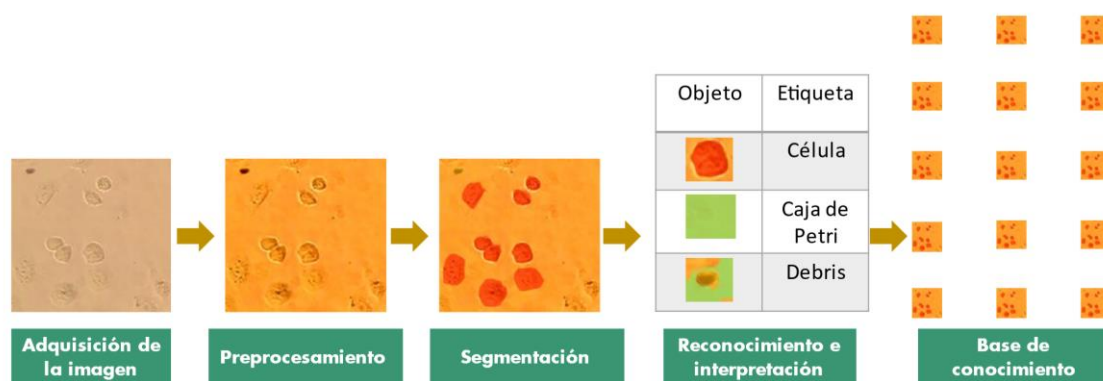


Figura 3: Esquema general de las etapas del procesamiento digital de imágenes.

2.3.1 Filtros espaciales para imágenes

Existen diversas técnicas que permiten mejorar la imagen, que son utilizadas con frecuencia con el propósito de resaltar características en específico y que sean de alguna manera mejor percibidas por el ojo humano. Las técnicas de procesamiento digital de imágenes pueden ser expresadas en dos dominios, estos son el de la frecuencia y el espacial. El dominio de la frecuencia está basado en la transformada de Fourier y utiliza filtros pasa-altas, pasa-bajas y pasa-bandas, por lo que se aplica con mayor frecuencia en la eliminación de ruido, además puede operar de forma discreta o continua. Por otro lado, el dominio espacial se basa en el plano de la imagen, manipulando los píxeles de forma directa, por lo que utiliza diversas técnicas de corrección de contraste, eliminación de ruido, detección de bordes y umbralizado [6].

Dependiendo de la aplicación y de la problemática a tratar, las técnicas mencionadas anteriormente pueden variar, por lo que, para el caso específico de las imágenes de líneas celulares utilizadas en el desarrollo de este trabajo de tesis, estas las técnicas filtrado espacial se usan para incrementar el contraste entre las regiones que contienen células de las que tienen residuos celulares u otros contaminantes de la muestra, llamados debris, y el del fondo de la caja de Petri.

Es decir, el objetivo de usarlos es resaltar los bordes de las células, sin superar cierto umbral, ya que la amplitud del fondo puede ocasionar ruido, lo que podría presentar una detección falsa de la célula. Dicho umbral es dependiente del enfoque y la intensidad de la luz con que fue capturada la imagen y la calidad del dispositivo con que fue tomada.

2.3.2 Filtro de difusión anisotrópica

El filtro de difusión anisotrópica es una técnica de procesamiento digital de imágenes que permite reducir el ruido de la imagen, mientras conserva el contorno de los objetos y suaviza el fondo de la imagen, además, realiza el filtrado en imágenes de 8 y 16 bits, al igual que en imágenes RGB [7].

En 1990 Malik y Perona presentaron un modelo de suavizado multiescala en imágenes con el planteamiento matemático relacionado con la percepción visual, en donde es necesaria la escala de la resolución, contemplando tres criterios, el suavizado según la zona, la localización inmediata y la causalidad [8]. El presente modelo brinda una mejora importante en el análisis multiescala para la detección de bordes, la ecuación diferencial que la describe es la siguiente:

$$\frac{\partial u}{\partial t} = \text{div}(g(|Du|Du)), u(0) = u_0 \quad (1)$$

En la ecuación (1), en donde $\text{div}()$ es el operador de divergencia, $g(Du)$ es el coeficiente de difusión. Sin embargo, esta ecuación no se encuentra propiamente en la dirección ortogonal al gradiente, por lo que existe la necesidad de crear una difusión en la dirección del vector:

$$\text{div}(g(|Du|Du)) = g(Du)\Delta u + g'(|Du|)|Du|^{-1}D^2u(Du, Du) \quad (2)$$

Considerando el termino de difusión en dirección ortogonal a Du y un término en dirección de Du :

$$u_{\xi\xi} = \frac{(u_{xx}u_y^2 - 2u_{xy}u_xu_y + u_{yy}u_x^2)}{|Du|^2} = D^2u\left(\frac{Du^\perp}{|Du^\perp|}, \frac{Du^\perp}{|Du^\perp|}\right) \quad (3)$$

$$u_{\eta\eta} = \frac{(u_{xx}u_x^2 - 2u_{xy}u_xu_y + u_{yy}u_y^2)}{|Du|^2} = D^2u\left(\frac{Du}{|Du|}, \frac{Du}{|Du|}\right) \quad (4)$$

Con ambas ecuaciones en dirección del vector Du y en dirección ortogonal a Du , la ecuación de difusión anisotrópica queda escrita de la siguiente forma:

$$\operatorname{div}(g(|Du|)Du) = g(Du)u_{\xi\xi} + G'(|Du|)u_{\eta\eta} \quad (5)$$

En la ecuación (5) si $G'(|Du|)$ es mayor a cero, la ecuación tenderá a mantener una difusión en ambas direcciones, en el caso contrario en donde $G'(|Du|)$ sea igual a cero, implicaría que la difusión es ortogonal al gradiente [7]. Como se puede apreciar en la forma en la que se desarrolló el modelo, este puede contener propiedades de invarianza, con lo que la ecuación podría ser escrita como:

$$\frac{\partial u}{\partial t} = F(D^2u, Du), \text{ en donde } F(A, p) = g'(|p|)|p|^{-1}A(p, p) + g(|p|) \quad (6)$$

Para la aplicación del filtro de forma práctica, se utilizan seis parámetros, estos se muestran y describen en la Tabla 1 [7]:

Tabla 1: Parámetros de filtro de difusión anisotrópica.

Parámetro	Descripción
Número de iteraciones	Cantidad máxima de iteraciones realizadas de forma completa.
Suavizados por iteración	Máscara Gaussiana de 3x3.
Limitador de difusión por variación mínima (A1)	Máscara de suavizado $F1(11,12) = (1 + 11 + 12)^{-A1}$, combinada linealmente con los pesos de los tensores F1 y F2, definidos por máscaras de 2x2.
Limitador de difusión por variación máxima (A2)	Máscara de suavizado $F2(11,12) = (1 + 11 + 12)^{-A2}$, combinada linealmente con los pesos de los tensores F1 y F2, definidos por máscaras de 2x2.
Paso del tiempo (Dt)	Paso de tiempo de una iteración.
Altura del umbral.	Define que tan marcados serán los bordes que conservará el filtro.



Figura 4: Efecto de filtrado de difusión anisotrópica en imágenes de células de cáncer de mama. a) Imagen original, b) imagen con 20 iteraciones del filtro (centro) y c) imagen con 60 iteraciones.

En la Figura 4, se muestra un ejemplo de la aplicación del filtro en una imagen de células de cáncer de mama, digitalizada a RGB de 603 x 573 píxeles, con dos cantidades distintas de iteraciones, en la que se puede apreciar como los bordes se conservan, pero el fondo de la imagen se ve más uniforme, sin embargo, existen limitaciones en la aplicación de este filtro, ya que para una imagen de 8 bits de 1.3MB, consume 220 MB de memoria RAM.

2.3.3 Filtro Kuwahara

El filtro Kuwahara es una técnica de procesamiento digital de imágenes que realiza un suavizado no lineal de la imagen, que permite mantener la integridad de los bordes, mientras elimina el ruido de la imagen, además, puede ser implementado con máscaras de distintas magnitudes, e inclusive de diferente forma [9]. En la Figura 5, se puede apreciar la aplicación del filtro Kuwahara en una imagen con distintas máscaras:

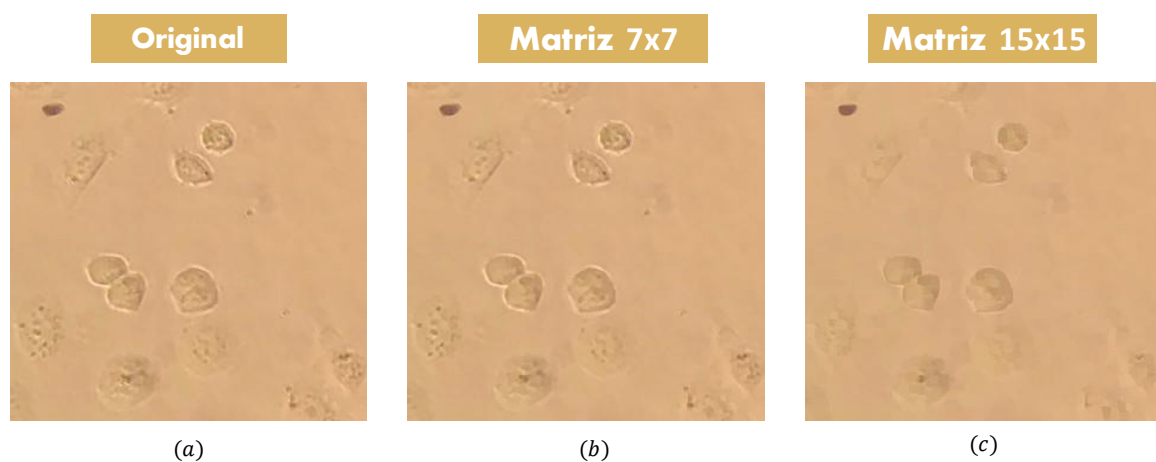


Figura 5: Efecto del uso del filtro Kuwahara (a) Imagen original, (b) imagen con máscara 7x7, (c) imagen con máscara 15x15.

La ecuación que modela la matriz del filtro es la siguiente:

$$H = W = 4L + 1 \quad (7)$$

En donde H es la altura de la máscara y W es la anchura de la misma, L es un número entero, y la constante 4 es referente a las regiones que mantendrá el enmascarado. A continuación, se muestra una máscara de 5x5:

$$\begin{bmatrix} a & a & \frac{a}{b} & b & b \\ a & a & \frac{a}{b} & b & b \\ \frac{a}{c} & \frac{a}{c} & \frac{\frac{a}{b}}{\frac{c}{d}} & \frac{b}{d} & \frac{b}{d} \\ c & c & \frac{c}{d} & d & d \\ c & c & \frac{c}{d} & d & d \end{bmatrix} \quad (8)$$

La máscara se encuentra dividida en cuatro regiones con el siguiente tamaño:

$$\frac{H+1}{2} \times \frac{W+1}{2} \quad (9)$$

Donde las cuatro regiones de la máscara son de tamaño 3x3, los pixeles que convergen entre dos regiones son pertenecientes a ambas regiones, es así que los pixeles se superponen, por lo que se calculan la desviación estándar y la media aritmética de las cuatro regiones y se utilizan para determinar el valor del píxel que se encuentra en el centro de la matriz. El resultado del píxel central será el promedio con relación a los píxeles convergentes [9].

2.3.4 Filtro de media

El filtro espacial de media, es una técnica de procesamiento digital de imágenes simple, ya que es fácil de implementar para el suavizado de imágenes, además, elimina el ruido, sin embargo, no mantiene los bordes y estos suelen perderse al aplicar el filtro de media. El filtro utiliza un enmascarado, con el que selecciona la media del píxel procesado en referencia a sus píxeles vecinos, la matriz puede ser vista de la siguiente forma en la Figura 6 y un ejemplo de la aplicación del filtro en la Figura 7, [10].

La máscara como la que se mostró previamente, no es más que un promedio aritmético ponderado alrededor del píxel central, dado por la siguiente ecuación:

$$g(i,j) = \frac{1}{n \times n} \sum_{i=1}^n \sum_{j=1}^n f(i,j) \quad (10)$$

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Figura 6: Máscara de 3x3 con filtro promedio.

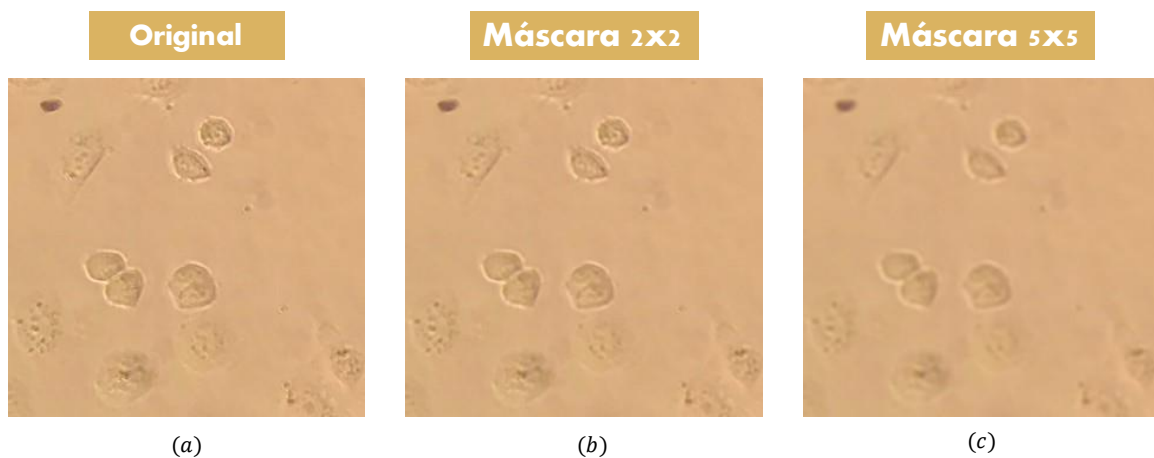


Figura 7: Efecto del uso del filtro de media (a) Imagen original, (b) imagen con máscara 2x2, (c) imagen con máscara 5x5.

En el que “n” es el tamaño del enmascarado, expresado por $\frac{1}{9}(a + b + c + d + e + f + g + h + i)$

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (11)$$

2.4 Procesamiento digital de imágenes usando FIJI®

La herramienta para el procesamiento de imágenes digitales utilizado en el presente trabajo de tesis es FIJI®, el cual es un software de código abierto y uso libre, que incluye diversos complementos para el análisis y mejora de imágenes. Puede ser utilizado por cualquier computadora con virtual Java 1.4 o posterior. FIJI® contiene el mismo menú que ImageJ2 [11], este puede observarse en la Figura 8.

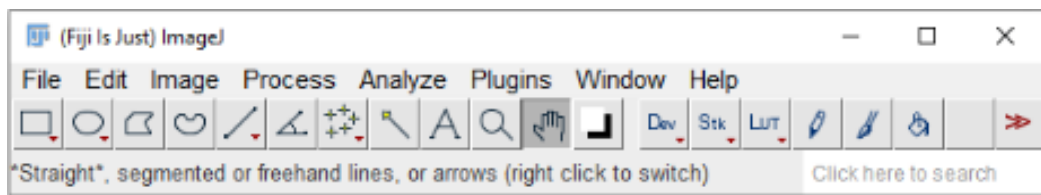


Figura 8: GUI de FIJI®, y su barra de menú.

Este menú, contiene distintos submenús que le permite al usuario editar, procesar, analizar, guardar y abrir imágenes de diferentes formatos: 8 bits, 16 bits, 32 bits, TIFF, GIF, JPEG, PNG, BMP DICOM y FITS, además, es un software multiproceso debido a que puede realizar operaciones en paralelo. Puede calcular valores estadísticos de áreas de la imagen y píxeles, se puede realizar mediciones de distancias entre píxeles en la imagen y ángulos sobre la misma. Se pueden crear histogramas de la intensidad de grupos de píxeles por componente de color o por tonos de gris, así como gráficos de perfil de línea. El usuario puede aplicar técnicas de preprocesado de imágenes como lo son: la corrección de contraste, el suavizado de imágenes, la eliminación de ruido y la detección de bordes.

El programa FIJI® admite crear y manejar simultáneamente una cantidad casi ilimitada de ventanas abiertas gracias a su arquitectura abierta, que proporciona extensibilidad a los complementos. Puede realizar transformaciones geométricas como la rotación o el cambio de escalas (de 32:1 a 1:32), además de, proporcionar una opción de calibración espacial con el propósito de permitir al usuario realizar mediciones dimensionales en unidades físicas de longitud como lo son los centímetros, como se puede apreciar es un software sumamente completo que permite desarrollar múltiples tareas, a continuación, se muestra el logotipo del software en la Figura 9.



Figura 9: Logotipo de software FIJI®.

2.4.1 Menú procesar

El menú Procesar (Process) es el cuarto menú de izquierda a derecha de la interfaz de usuario del programa FIJI®. Este menú incluye los comandos relacionados con el procesamiento de imágenes digitales, dentro de las cuales se incluyen los filtros, las operaciones puntuales y las operaciones aritméticas con imágenes [12], estos son vistos en la Figura 10. Para efectos de la aplicación de los comandos más relevantes para el presente trabajo de tesis, solo se hará mención de los siguientes:

- Smooth (Suavizado).
- Sharpen (Agudizado).
- Find edges (Detector de bordes).
- Find máxima (Máximos locales).
- Enhance contrast (Corrección de contraste).

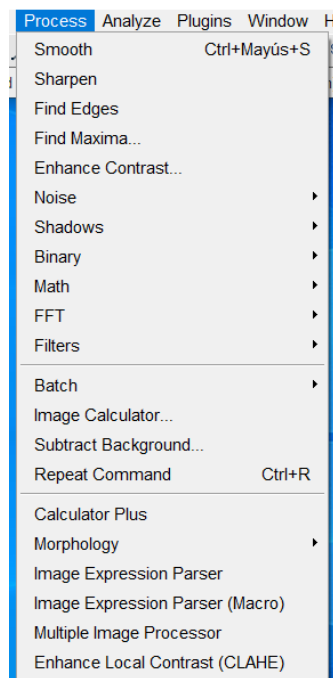


Figura 10: Menú procesar del software FIJI®.

El primer comando para preprocesar digitalmente las imágenes es Smooth (Suavizar), emplea un filtro de suavizado que reemplaza cada píxel con el promedio de sus adyacentes, usa un enmascarado de una matriz de 3x3. El efecto de la aplicación del filtro de suavizado en una imagen de células de cáncer de mama se muestra en la Figura 11.

El filtro de Agudizado (Sharpen) es el siguiente comando de preprocesado en el menú, este incrementa el contraste y acentúa los detalles de la imagen, utiliza factores de ponderación para reemplazar cada píxel con un promedio ponderado de sus adjuntos, usa un enmascarado 3x3. La matriz correspondiente a dicha máscara se muestra en la Figura 12. Una desventaja de este filtro es que también puede agudizar el ruido de la imagen. Un ejemplo del efecto de filtro de Agudizado (Sharpen) en una imagen de células de cáncer de mama puede ser visto en la Figura 13, [12].

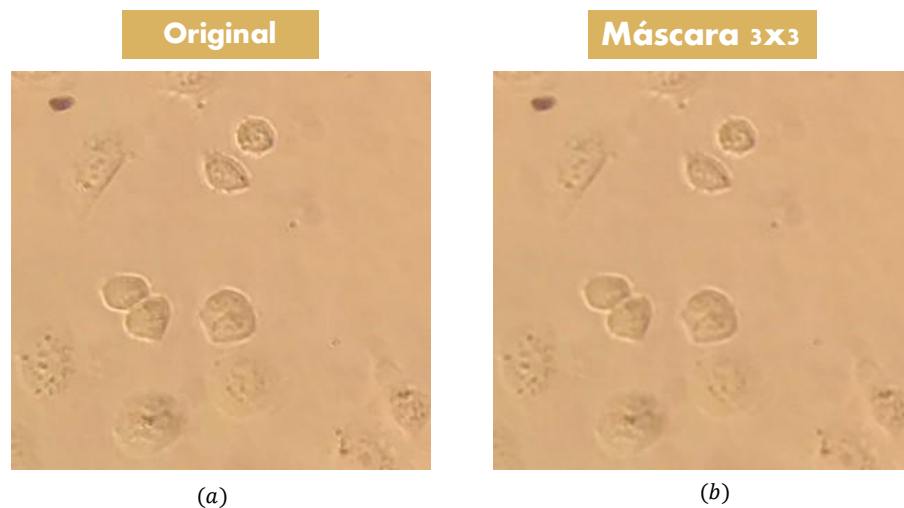


Figura 11: Efecto de la aplicación del filtro de suavizado en una imagen de células de cáncer de mama. (a) Imagen original, (b) imagen con filtro de suavizado (Smooth).

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 12 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Figura 12: Máscara 3x3 de filtro de agudizado.

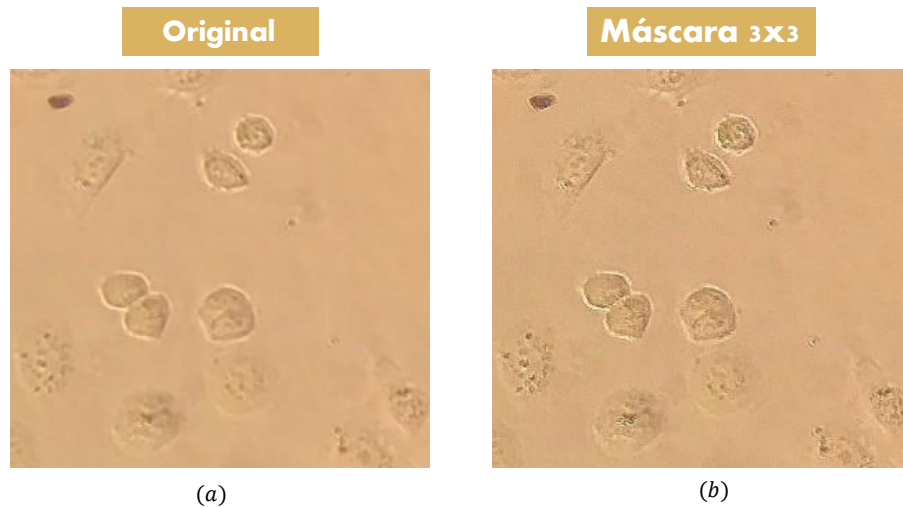


Figura 13: Efecto de filtro de Agudizado (Sharpen) en una imagen de células de cáncer de mama. Imagen original (a), (b) imagen con filtro de Agudizado (Sharpen).

El detector de bordes (Find Edges) utiliza el operador gradiente de Sobel para resaltar los cambios abruptos de intensidad en la imagen, se basa en dos máscaras convolucionadas de 3x3, mostradas en la Figura 14, con las cuales generan una combinación de las derivadas verticales y horizontales, mediante la magnitud de las mismas [12]. A continuación, se puede observar el efecto de aplicar el detector de bordes (Find Edges) en una imagen de cáncer de mama, mediante la aplicación del operador gradiente en la Figura 15.

$$\begin{array}{ccc|ccc}
 1 & 2 & 1 & 1 & 0 & -1 \\
 0 & 0 & 0 & 2 & 0 & - \\
 -1 & -2 & -1 & 1 & 0 & -1
 \end{array}$$

Figura 14: Máscara 3x3 de barrido horizontal (izquierda), máscara 3x3 de barrido vertical (derecha).

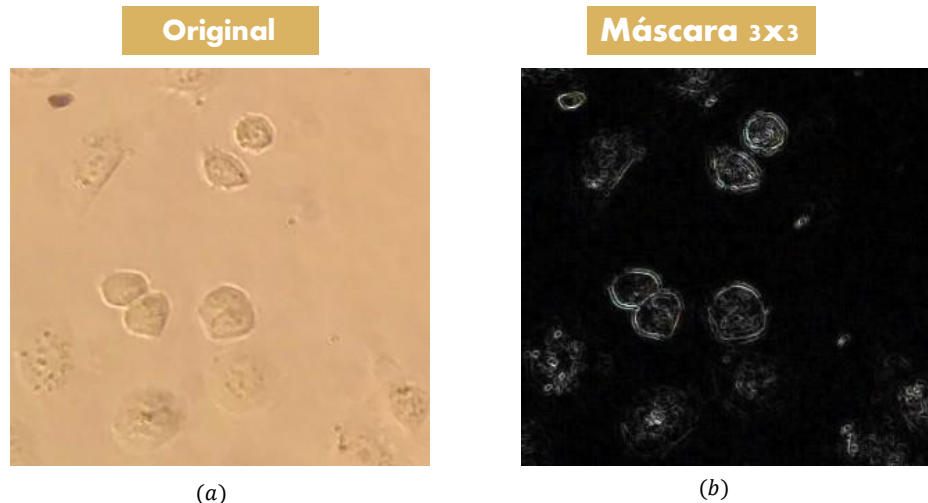


Figura 15: Efecto de aplicar el detector de bordes (Find Edges) en una imagen de cáncer de mama Imagen original (a), imagen con detector de bordes (Find edges) del menú procesar (b).

Máxima (Find maxima) determina los máximos locales en una imagen, tomando como referencia la luminancia y su promedio ponderado, creando una máscara del mismo tamaño con los máximos, el análisis es realizado en una sección determinada o en la imagen completa, en la Figura 16 se puede apreciar cómo es aplicado, además, en la Figura 17 se observan los valores obtenidos de los parámetros [12].

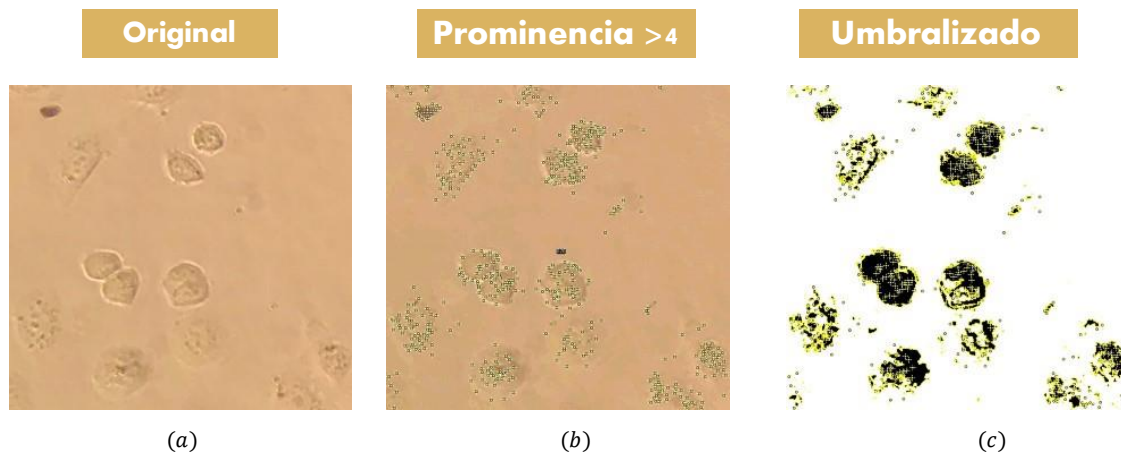


Figura 16: Efecto de aplicar el detector de máximos locales (Find maxima) en una imagen de cáncer de mama Imagen original (a), imagen con máximos locales (Find maxima) (b), Figura umbralizado con los máximos locales (c).

Results

File Edit Font Results																						
	Label	Area	X	Y	Perim.	Major	Minor	Angle	Circ.	Feret	IntDen	Median	%Area	RawIntDen	Slice	FeretX	FeretY	FeretAngle	MinFeret	AR	Round	Solidity
1	IMG_9415_2_3_find maxima_umbralizado.jpg 3	10.833	2.167	5.657	2.365	1.615	135.000	1.000	2.828	177	53	100	177	1	10	1	135	2	1.464	0.683	0.857	
2	IMG_9415_2_3_find maxima_umbralizado.jpg 2	38.000	1.500	4.243	2.257	1.128	0.000	1.000	2.236	154	77	100	154	1	37	1	153.435	1	2.000	0.500	1.000	
3	IMG_9415_2_3_find maxima_umbralizado.jpg 1	73.500	2.500	2.828	1.128	1.128	0.000	1.000	1.414	53	53	100	53	1	73	2	135.000	1	1.000	1.000	1.000	
4	IMG_9415_2_3_find maxima_umbralizado.jpg 259	267.396	12.280	142.995	25.186	13.094	6.057	0.159	35.355	21971	70	100	21971	1	253	11	171.870	20.823	1.924	0.520	0.497	
5	IMG_9415_2_3_find maxima_umbralizado.jpg 1	35.500	3.500	2.828	1.128	1.128	0.000	1.000	1.414	77	77	100	77	1	35	3	135.000	1.000	1.000	1.000	1.000	
6	IMG_9415_2_3_find maxima_umbralizado.jpg 16	242.625	6.375	22.385	5.706	3.570	26.643	0.401	7.616	1088	68	100	1088	1	239	8	23.199	5.578	1.598	0.626	0.533	
7	IMG_9415_2_3_find maxima_umbralizado.jpg 34	33.941	7.559	36.042	12.688	3.412	17.798	0.329	13.928	2406	68	100	2406	1	28	10	21.038	4.743	3.719	0.269	0.630	
8	IMG_9415_2_3_find maxima_umbralizado.jpg 1	29.500	5.500	2.828	1.128	1.128	0.000	1.000	1.414	77	77	100	77	1	29	5	135.000	1.000	1.000	1.000	1.000	
9	IMG_9415_2_3_find maxima_umbralizado.jpg 32	46.938	8.094	34.385	10.800	3.773	16.077	0.340	13.928	2244	68	100	2244	1	40	11	21.038	5.000	2.863	0.349	0.634	
10	IMG_9415_2_3_find maxima_umbralizado.jpg 4	235.500	7.250	7.071	2.734	1.863	0.000	1.000	3.162	212	53	100	212	1	234	7	161.565	2.000	1.468	0.681	0.800	
11	IMG_9415_2_3_find maxima_umbralizado.jpg 10	263.100	8.100	13.899	4.708	2.705	17.877	0.650	5.831	714	71	100	714	1	260	9	30.964	3.578	1.740	0.575	0.714	
12	IMG_9415_2_3_find maxima_umbralizado.jpg 21	484.738	8.024	19.556	7.099	3.766	11.809	0.690	8.544	1606	68	100	1606	1	481	9	20.556	4.000	1.885	0.531	0.792	
13	IMG_9415_2_3_find maxima_umbralizado.jpg 1	296.500	7.500	2.828	1.128	1.128	0.000	1.000	1.414	68	68	100	68	1	296	7	135.000	1.000	1.000	1.000	1.000	
14	IMG_9415_2_3_find maxima_umbralizado.jpg 1	301.500	7.500	2.828	1.128	1.128	0.000	1.000	1.414	68	68	100	68	1	301	7	135.000	1.000	1.000	1.000	1.000	
15	IMG_9415_2_3_find maxima_umbralizado.jpg 2	13.000	8.500	4.243	2.257	1.128	0.000	1.000	2.236	136	68	100	136	1	12	8	153.435	1.000	2.000	0.500	1.000	
16	IMG_9415_2_3_find maxima_umbralizado.jpg 8	70.125	8.750	8.889	4.044	2.519	160.670	1.000	4.472	424	53	100	424	1	68	8	153.435	2.000	1.606	0.628	0.800	

Figura 17: Tabulación de parámetros de análisis de imágenes.

El quinto comando en el menú de preprocesado es la corrección de contraste (enhance contrast), esto se realiza mediante la ecualización del histograma de la imagen. En este procedimiento no se alteran los valores de los píxeles de la imagen, solo se realiza un cambio en la saturación de los píxeles. Al aumentar el valor de la saturación se aumenta proporcionalmente el contraste, esto se hace cuidando que el histograma no se extienda debido a valores de la saturación cercanos a cero. El efecto de aplicar la corrección de contraste (enhance contrast) en una imagen de células cáncer de mama se muestra en la Figura 18, [12].

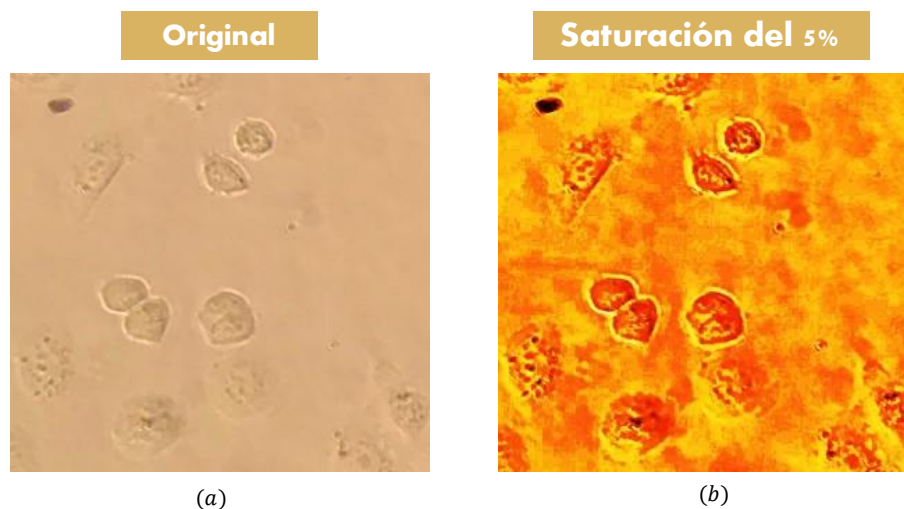


Figura 18: Efecto de aplicar la corrección de contraste (enhance contrast) en una imagen de células cáncer de mama. (a) Imagen original, (b) imagen con corrección de contraste (Enhance contrast).

2.4.2 Menú analizar

El menú Analizar (Analyze) es el quinto menú de la interfaz de usuario del software FIJI®, como se muestra en la Figura 19, este contiene comandos relacionados con mediciones estadísticas de datos de imágenes, trazado de histogramas y complementos para el análisis de imágenes [12]. Para efectos de la aplicación de los comandos más relevantes para el presente trabajo de tesis, solo se hará mención de los siguientes:

- Analizar partículas (Analyze particles).
- Establecer medidas (Set measurements).

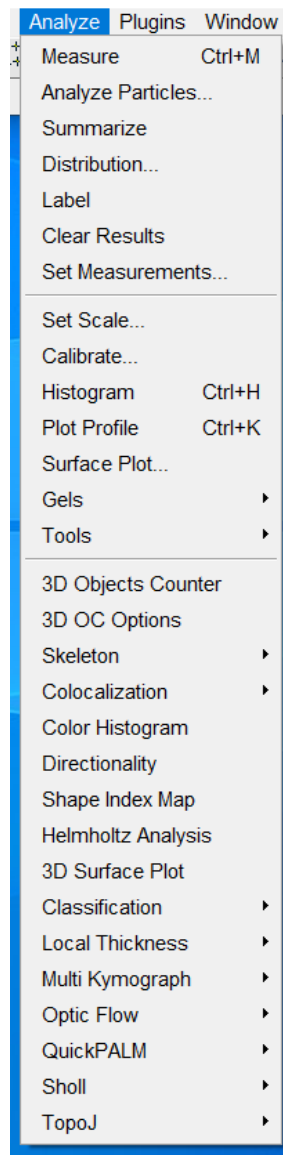


Figura 19: Menú analizar del software FIJI®.

El submenú Analizar partículas (Particle Analyze) cuenta y mide objetos en imágenes con formato binario, realiza su función escaneando la imagen en búsqueda de bordes. El acto de detectar los bordes de un objeto implicaría que existe el objeto. Este se delinea para posteriormente ser rellenado y medido, esto se puede aplicar en el área seleccionada o en toda la imagen. Contiene tres parámetros que le permiten al usuario indicar el área de los objetos a analizar y la clase de máscara que se requiere a la salida del proceso, estos parámetros son los siguientes:

- Tamaño (Size).
- Circularidad (Circularity).
- Mostrar (Show).

El parámetro de Tamaño (Size) especifica el rango del área que se tomara en cuenta para el proceso, el valor debe encontrarse entre 0 e ∞ . Este valor es expresado en unidades de superficie o en píxeles. Debido a esto, si el usuario ingresa un rango de valores y las partículas de la imagen son menores o mayores al rango especificado, estas serán ignoradas [12].

En el parámetro Circularidad (Circularity) las partículas con valores de tamaño fuera del rango especificado serán ignoradas al igual que en el parámetro anterior. La circularidad es dada por la ecuación (12), en la que el rango se encuentra de 0 (polígono con caras infinitas) a 1 (circulo perfecto). En la Figura 20, se muestra el cambio el efecto de aplicar el análisis de partículas (enhance contrast) en una imagen de células cáncer de mama en base a los parámetros tamaño y la circularidad [12].

$$4\pi \frac{\text{área}}{(\text{perímetro})^2} \quad (12)$$

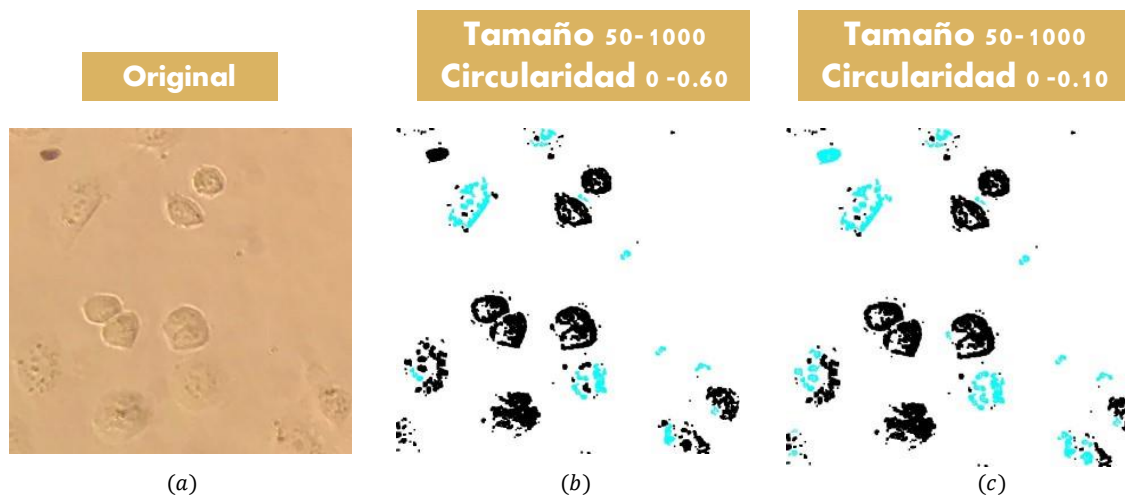


Figura 20: Efecto de aplicar el análisis de partículas (Enhance Contrast) en una imagen de células de cáncer de mama en base a los parámetros tamaño y la circularidad. (a) Imagen original, (b) imagen analizada con tamaño 50-1000 y circularidad 0-0.60, (c) imagen analizada con tamaño 50-1000 y circularidad 0-0.10.

El parámetro **Mostrar** especifica la máscara, imagen o superposición que debe mostrarse después del análisis. Contiene los siguientes tipos de enmascarado:

- Contornos.
- Contornos expuestos.
- Contornos superpuestos.
- Máscaras.
- Máscaras superpuestas.

El siguiente comando permite es el de **Establecer Medidas** (set measurements), este le permite al usuario establecer las características que se extraerán de la imagen o máscara, con base a dos grupos de parámetros, el primer grupo selecciona las mediciones, mientras que el segundo grupo brinda un ajuste a las misma. En la Figura 21 se puede visualizar el submenú para establecer dichas medidas [12].

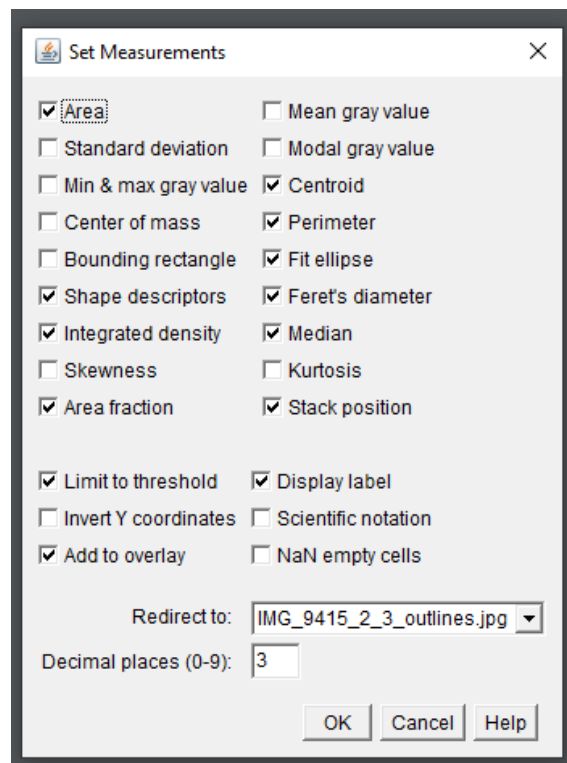


Figura 21: Submenú establece medidas del software FIJI®.

En el mencionado menú se muestran dos grupos de parámetros, los cuales se mencionan a continuación junto con su descripción. El primer grupo de parámetros esta está formado por:

- **Área:** Permite al usuario calibrar la imagen espacialmente en píxeles o en milímetros.
- **Desviación estándar:** Permite al usuario obtener la desviación estándar de los valores de gris en la imagen.
- **Valores máximos y mínimos en escala de grises:** Indica al usuario los valores máximos y mínimos de gris de una sección de píxeles o de toda la imagen,
- **Centro de masa:** Le permite al usuario obtener información acerca del promedio ponderado del brillo de los píxeles, indicando los momentos espaciales de primer orden.
- **Descriptor de Figuras:** Le brinda al usuario información acerca de los parámetros circularidad, solidez, relación de aspecto y redondez.
- **Densidad integrada:** Muestra al usuario el producto del área y el valor de la media en escala de grises y la suma de los valores de los píxeles en la sección designada o de la imagen completa.
- **Asimetría:** Indica al usuario el momento de tercer orden con respecto a la media.
- **Área fraccionada:** Indica al usuario el porcentaje de píxeles que se han resaltado en la sección designada o en la imagen completa.
- **Valor de la media en escala de grises:** Brinda al usuario la suma de la media de los valores de gris.
- **Valor de la moda en escala de grises:** Brinda al usuario el valor de gris que se encuentre con más frecuencia, lo que en otras palabras es el valor más alto de un histograma.
- **Centroide:** Permite al usuario obtener el promedio de las coordenadas (x, y), utilizando los encabezados de la Tabla de resultados.
- **Perímetro:** Brinda al usuario la longitud del límite exterior de la imagen o de la sección que se designó.
- **Ajuste de elipse:** Le permite al usuario ajustar la elipse que se marca con el parámetro centroide, utilizando los ejes primarios y secundarios de la elipse, y el ángulo entre el eje principal y la línea paralela al eje “x”.

- **Diámetro de Feret:** También conocido como calibre máximo, es la máxima distancia entre dos puntos a lo largo del límite de una sección. Brinda al usuario información acerca de los parámetros Feret, ángulo de Feret, mínimo de Feret, y las coordenadas X y Y de Feret.
- **Mediana:** Le indica al usuario el valor de la mediana de los píxeles de la imagen o sección designada.
- **Curtosis:** Le brinda al usuario información acerca del grado de concentración que mantienen los datos de una variable con respecto a la zona central de la gráfica, para este caso en específico indica el momento de cuarto orden sobre la media.
- **Posición en la pila:** Permite al usuario agregar a la pila la posición actual (segmento, marco o canal) [12].

El segundo grupo está formado por los siguientes parámetros, sus nombre y descripciones se proporcionan a continuación:

- **Umbral limitado:** Permite al usuario incluir en los cálculos de medición a los píxeles seleccionados en el umbral.
- **Coordenadas “Y” invertidas:** Permite al usuario invertir el origen del plano, ubicándolo en la esquina inferior izquierda.
- **Agregar enmascarado:** Permite al usuario visualizar el enmascarado en otra imagen.
- **Visualizar etiqueta:** Permite al usuario indicar el nombre de la imagen y el número de corte.
- **Notación científica:** Permite al usuario visualizar los valores con notación científica [12].

2.4.3 Menú complementos.

El menú complementos es el sexto menú de la interfaz de usuario del software FIJI® y se muestra en la Figura 22, este contiene comandos relacionados con la escritura de macros con que el software dispone para el análisis de imágenes, además, FIJI® contiene complementos [12]. Para efectos de la aplicación de los comandos más relevantes para el presente trabajo de tesis, solo se hará mención del siguiente:

- Segmentación Entrenable de Weka (Trainable Weka Segmentation).

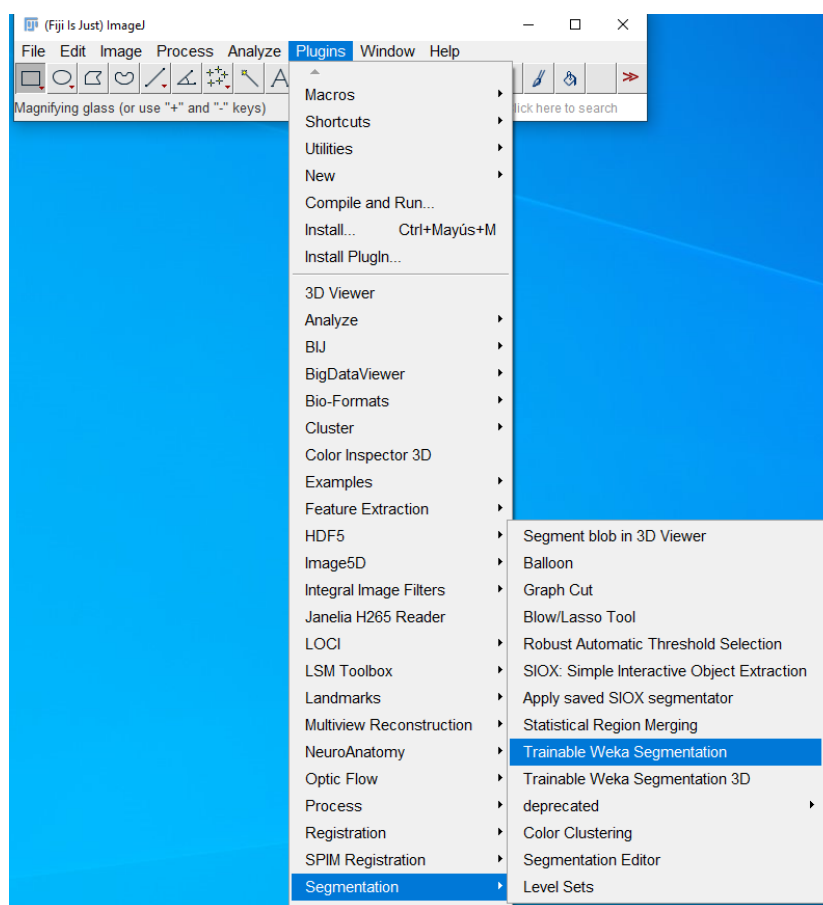


Figura 22: Menú complementos del software FIJI®.

La opción de Segmentación Entrenable de Weka (TWS) del software FIJI®, es un complemento que le permite al usuario utilizar en conjunto algoritmos de inteligencia artificial del software WEKA® y técnicas de procesamiento digital de imágenes por parte del software FIJI®. Este complemento puede entrenarse con el objetivo de aprender de la mano del usuario y realizar la tarea designada con datos de prueba, además, utiliza los algoritmos de aprendizaje automatizado de WEKA®, con lo que proporciona al usuario imágenes etiquetadas, con el entrenamiento del clasificador seleccionado. La opción TWS segmenta las imágenes clasificando píxeles, el proceso en la segmentación de este complemento se puede apreciar en la Figura 23.

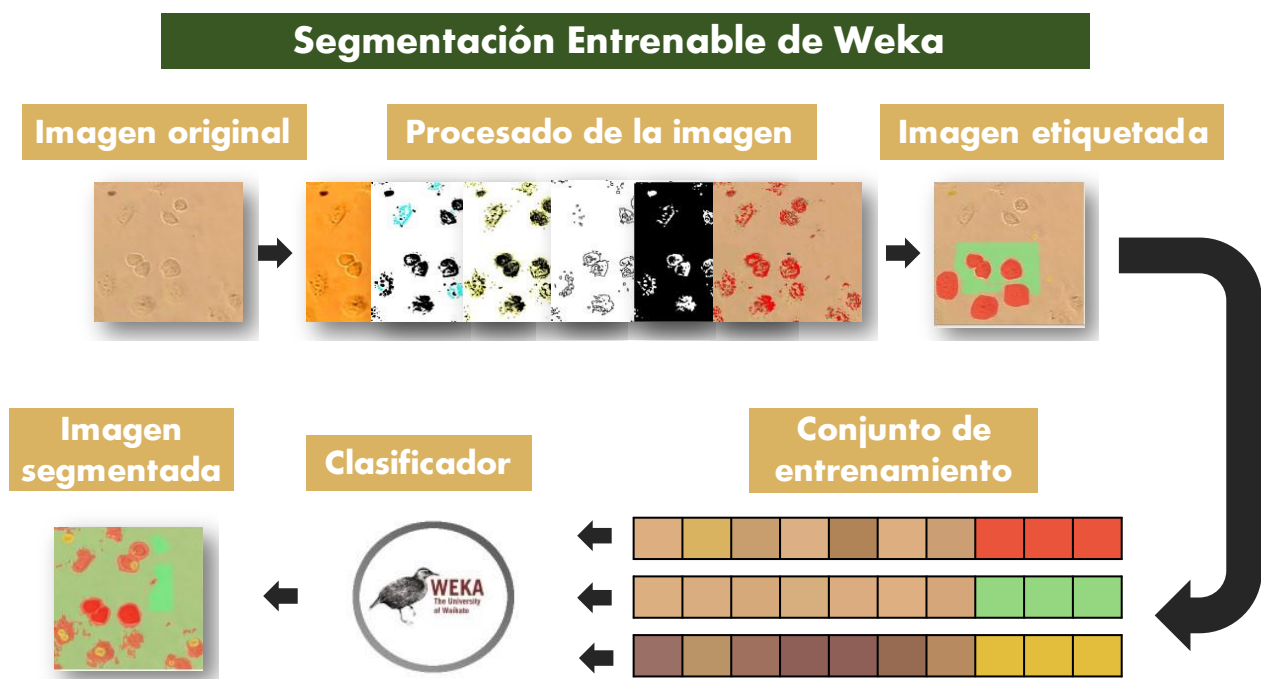


Figura 23: Proceso de Segmentación Entrenable de Weka (TWS).

El proceso de segmentación con base a la opción TWS consta de cinco pasos de preprocesado, presentados anteriormente. Comienza con la adquisición de la imagen original, continúa con el procesado y extracción de características de la imagen, posteriormente la imagen es etiquetada para entrenar al algoritmo seleccionado, y finalmente se clasifica la imagen con uno de los algoritmos de aprendizaje supervisado de WEKA®. Este proceso se muestra en la Figura 23.

La interfaz gráfica de usuario del complemento TWS se muestra en la Figura 24, esta permite ejecutar imágenes en dos dimensiones (escala de grises) o en tercera dimensión (a color), en la Figura 24 se visualiza la interfaz gráfica. Al inicializar el TWS, se predefinen dos clases, sin embargo, el usuario puede agregar las clases que considere necesarias para el desarrollo de su trabajo. Esto se hace utilizando el push button Create New Class.

Un elemento de una clase se define utilizando el conjunto de herramientas para el dibujo. Para ello se selecciona de la región de interés en la imagen y se selecciona el push button con el nombre de la clase a la que pertenece para ser agregado con su etiqueta correspondiente. Este procedimiento debe ser realizado bajo una atenta supervisión de un experto en el área, puesto que el resultado de este proceso es la referencia para el proceso de aprendizaje. Las clases y el etiquetado realizado durante el proceso pueden ser apreciadas en la sección de etiquetas de la interfaz gráfica.

El panel de control de TWS contiene dos menús, el primero es utilizado para el entrenamiento y el segundo para aplicar los algoritmos y configurar los parámetros y características [13]. Los comandos pertenecientes al primer menú son los siguientes:

- **Entrenar clasificador (Train classifier):** Este comando inicializa el entrenamiento del algoritmo de clasificación seleccionado, extrae las características de la imagen de entrada, con lo que convierte la imagen en un conjunto de vectores. Al finalizar el entrenamiento, la imagen de salida será una superposición de la imagen original con la segmentación de las clases en un enmascarado.
- **Alternar superposición (Toggle overlay):** Este comando superpone la imagen de salida sobre la imagen de entrada.

- **Crear resultado (Create result):** Este comando muestra la imagen resultante sin superposición, en un formato de 8 bits, conservando los colores de las clases designadas.
- **Obtener probabilidad (Get probability):** El presente comando indica la probabilidad de cada píxel de pertenecer a una clase, tomando como referencia una hiperpila de 32 bits.
- **Graficar resultado (Plot result):** Este comando permite visualizar los gráficos de desempeño del algoritmo, en función de la precisión y recuperación basada en el entrenamiento [14].

Los comandos pertenecientes al segundo menú son:

- **Aplicar clasificador (Apply classifier):** Este comando permite aplicar el clasificador seleccionado en una imagen o en un grupo de imágenes, obteniendo como resultado imagen o imágenes segmentadas o un mapa de probabilidad.
- **Cargar clasificador (Load classifier):** Desde este comando se le permite al usuario cargar cualquier clasificador guardado en el complemento de TWS previamente, el formato utilizado es “. model”.
- **Guardar clasificador (Save classifier):** Este comando permite al usuario almacenar el clasificador, con el objetivo de aplicarlo posteriormente, el formato utilizado es “. model”.
- **Cargar datos (Load data):** Desde este comando el usuario tiene la facultad de cargar los datos guardados previamente, el formato utilizado es “. arff”.
- **Salvar datos (Save data):** Este comando permite al usuario almacenar los datos (vectores de características), con el objetivo de aplicarlo posteriormente, el formato utilizado es “. arff”.
- **Crear nueva clase (Create new class):** Este comando permite crear nuevas clases.
- **Ajustes (Settings):** Desde este comando es posible ajustar los parámetros, con los que se crearan los vectores de características, la ventana de este comando se puede apreciar en la Figura 26 [14].



Figura 24: Interfaz gráfica de usuario de Segmentación Entrenable de Weka (TWS).

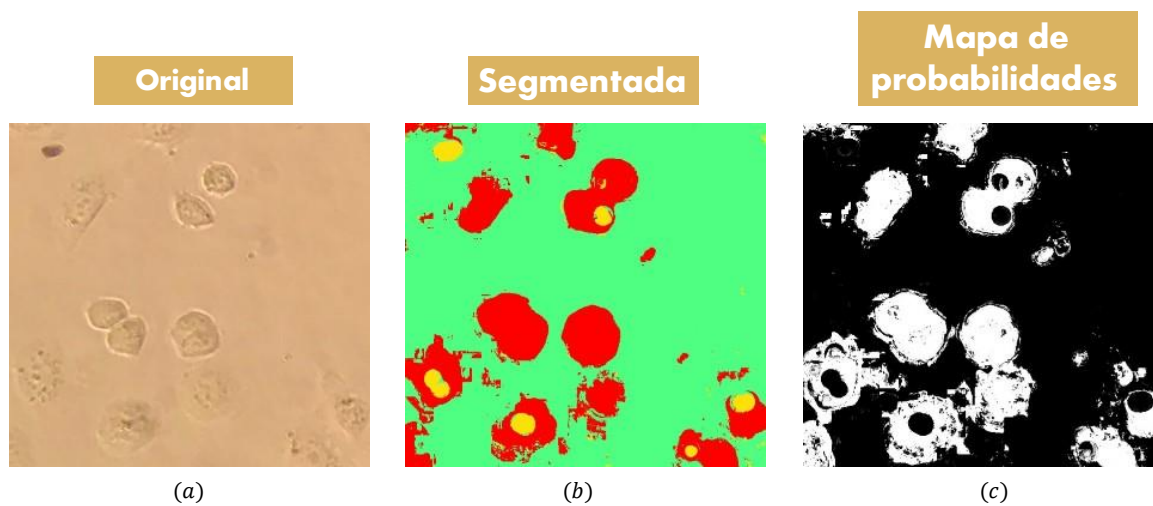


Figura 25: imagen original (a), imágenes resultantes de Segmentación Entrenable de Weka (TWS), la imagen segmentada (b) y el mapa de probabilidades de la imagen (c).

Se pueden observar imágenes resultantes de TWS en la Figura 26.

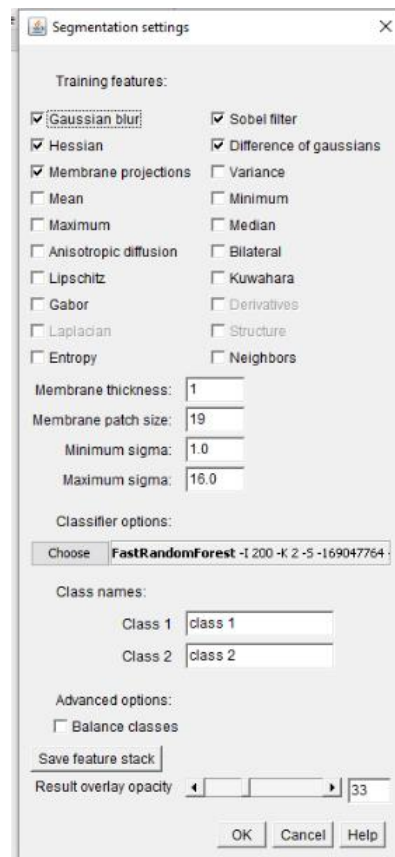


Figura 26: Ventana de ajustes de configuración de TWS

2.5 Algoritmos de Inteligencia Artificial para Imágenes

La inteligencia artificial (I.A.) comenzó a desarrollarse en la década de los 50's, de la mano del profesor John McCarthy, quien construyó un lenguaje de programación que lograba adaptarse a la reproducción de listas, con el objetivo de simular las actividades intelectuales que el hombre empleaba, imitando por medio de dispositivos las acciones mentales del ser humano [15]. En la literatura especializada se encuentran diversas definiciones de inteligencia artificial, sin embargo, todas llegan al mismo propósito designado para la inteligencia artificial, y este es el análisis, comprensión y percepción de problemáticas y su respectiva toma de decisiones, enfocándose en la simulación del análisis humano. Sin embargo, no existe una definición exacta, debido a que la inteligencia artificial se encuentra en constante cambio y se desarrolla en diversos ámbitos, por lo que solo se pueden tener aproximaciones a la definición de esta.

La inteligencia artificial está formada por un conjunto de algoritmos que tienen la capacidad de ser entrenados con el propósito de analizar y tomar decisiones en base a preceptos. La inteligencia artificial es una tecnología emergente con diversas aplicaciones, en múltiples campos de la ciencia y la ingeniería, utilizada ampliamente en la clasificación y regresión de datos. Existen dos vertientes en los algoritmos de inteligencia artificial, los algoritmos supervisados y los no supervisados.

Los algoritmos supervisados aprenden con base a un conjunto de datos que son etiquetados previamente, por lo que la salida del sistema es conocida. El objetivo de este tipo de inteligencia artificial es predecir atributos ya sean categóricos o numéricos. Por otro lado, los algoritmos no supervisados, funcionan con conjuntos de datos sin un etiquetado previo o clases definidas con anterioridad, sin embargo, son algoritmos que emplean tareas de agrupamiento, cuyo objetivo es encontrar similitudes en los conjuntos de datos [16]. A continuación, en la Tabla 2, se muestran los algoritmos clasificados según su tipo:

Tabla 2: Modelos de aprendizaje supervisado (izquierda) y modelos de aprendizaje no supervisados (derecha).

Algoritmos supervisados	Algoritmos no supervisados
Regresión por mínimos cuadrados.	Análisis de componentes principales.
Arboles de decisión.	Análisis de componentes independientes.
Máquinas de soporte vectorial.	Descomposición por valores singulares.
Impulso adaptativo.	K-means.
Naive Bayes.	Agrupamiento espacial basado en la densidad de aplicaciones con ruido.
Bosque aleatorio.	Propagación de afinidad.
Vecinos más cercanos.	Mezclas Gaussianas.
Redes neuronales	Incrustación de vecinos estocásticos distribuidos en T.

2.5.1 Etapas de la aplicación de algoritmos de inteligencia artificial

Generalmente los algoritmos de inteligencia artificial suelen seguir cierto patrón para lograr definir una agrupación, clasificación o regresión. En el caso específico de los algoritmos supervisados, es necesario que un experto determine el procesamiento de los datos, las características a extraer con la intención de definir atributos, el etiquetado de los datos, la partición del conjunto de datos de entrenamiento y de prueba, además, es necesario configurar los parámetros del algoritmo seleccionado. El proceso de aplicación de los algoritmos está representado en la siguiente Figura 27:

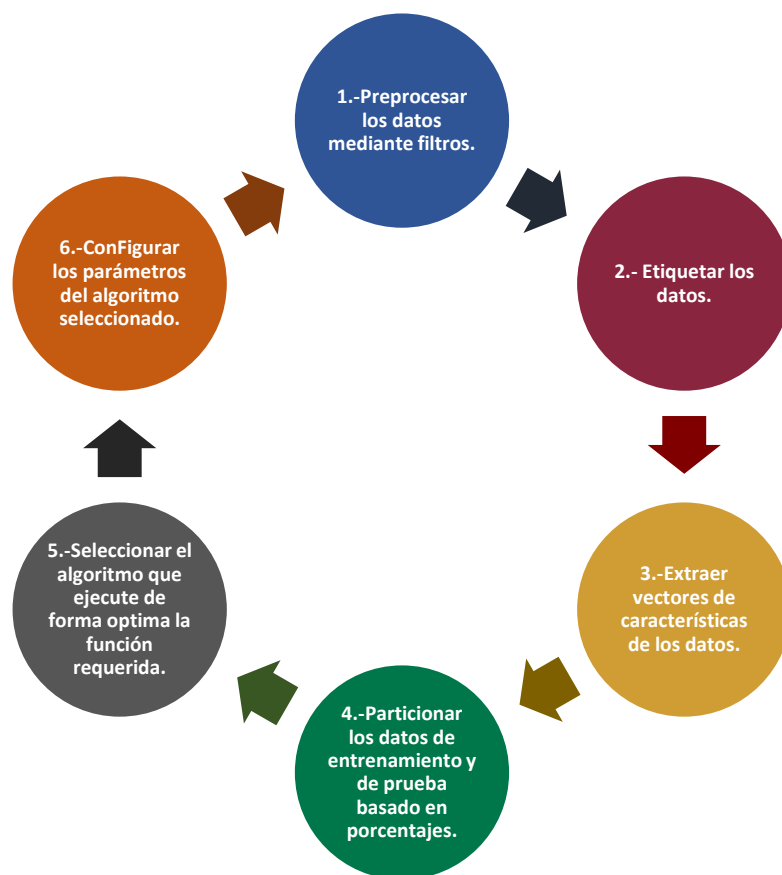


Figura 27: Proceso general en la aplicación de algoritmos de inteligencia artificial.

2.6 Naive Bayes

Thomas Bayes el matemático inglés que postuló su teorema en 1763, fue la inspiración para que en la década de los 50's se aplicará un algoritmo computacional basado en su teorema probabilístico. El algoritmo de Naive Bayes es un algoritmo de inteligencia artificial con un enfoque probabilístico simple, se basa en el teorema de Bayes, implicando que existe una independencia entre las características o predictores. El algoritmo estima el valor de los parámetros de media y varianza de las variables del conjunto de datos de entrenamiento, por lo que solo necesita una estimación de cada variable, sin la necesidad de generar una matriz de covarianza. Debido a que Naive Bayes presenta con frecuencia una alta precisión para conjuntos de datos en tiempo real, y que, además, requiere de una cantidad mínima de datos de entrenamiento, es uno de los algoritmos más utilizados [17].

La estimación de los parámetros del algoritmo de Naive Bayes está dada por la máxima verosimilitud, sin embargo, la probabilidad provista por este algoritmo es condicional. Esto es debido a que las variables categóricas independientes contienen diversas categorías y estas dependen de los predictores, por lo que, sí la cantidad de características es grande, el modelo debe ser modificado. La probabilidad de cada atributo se obtiene de forma independiente. El algoritmo predice rápidamente la clase de los datos de prueba, sin embargo, una limitación que sufre Naive Bayes es la nula posibilidad de que exista algún predictor independiente en un conjunto de datos real [18].

Una opción para analizar el desempeño del algoritmo de Naive Bayes es a través de la representación gráfica, en la cual se pueden apreciar las regiones en las que las clases se posicionan. En la Figura 28 se muestran los resultados de aplicar el algoritmo de Naive Bayes al conjunto de datos "Iris". En la Figura se puede observar que el algoritmo discrimina correctamente los datos, existiendo una separación clara entre las clases, sin confusiones. Esto muestra que el modelo de aprendizaje se desempeña eficientemente.

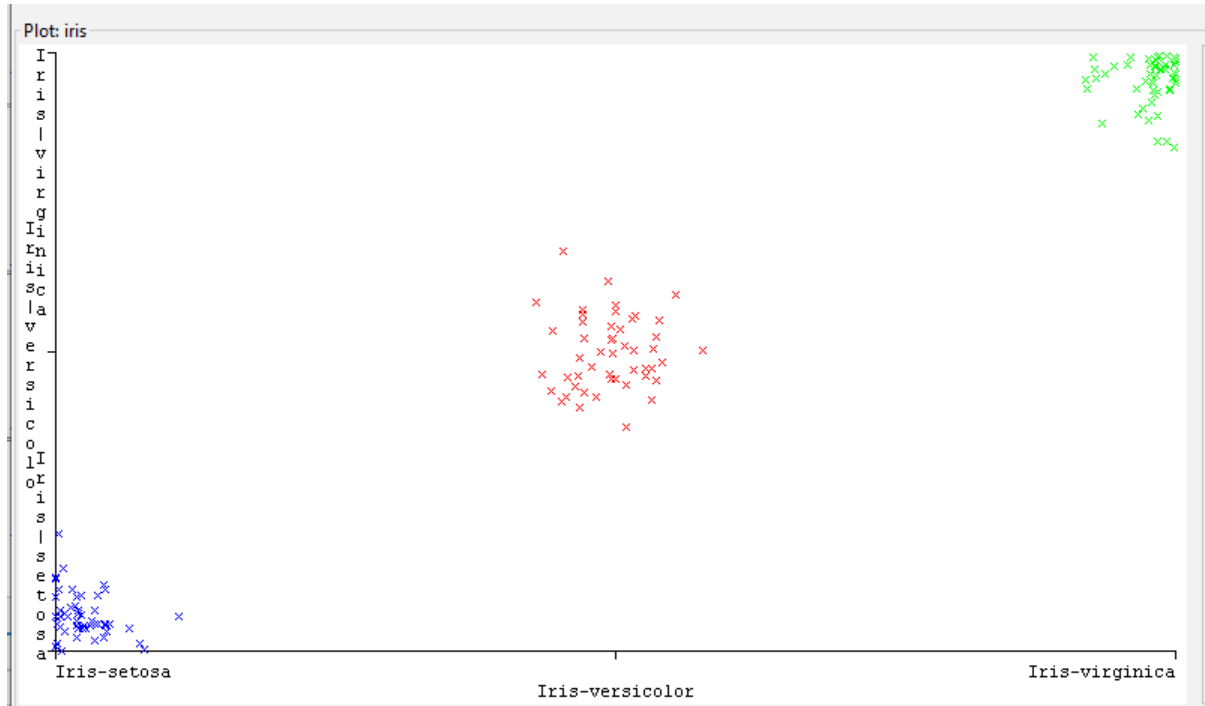


Figura 28: Resultados de aplicar el algoritmo de Naive Bayes al conjunto de datos “Iris”

2.6.1 Aprendizaje de Naive Bayes

El algoritmo de inteligencia artificial de Naive Bayes es un modelo supervisado, que utiliza el teorema de Bayes, el modelo supone que los atributos son condicionalmente independientes en referencia al valor dado a la clase. Los atributos predicen el valor de la clase, y estos son representados con $A_1, A_2, A_3 \dots A_n$, mientras que la clase es con " C "[18]. La función probabilística que define al presente algoritmo es la siguiente:

$$P(C = c/a_1, a_2, a_3 \dots a_n) = \frac{P(C=c)*P(a_1, a_2, a_3 \dots a_n / C=c)}{P(a_1, a_2, a_3 \dots a_n)} \quad (13)$$

En donde:

- $a_1, a_2, a_3 \dots a_n$ son un conjunto de atributos.
- C es una clase determinada.
- c es una etiqueta dada.

Lo que, aplicado al algoritmo, en términos de las características y etiquetas del conjunto de datos se apreciaría de la siguiente forma:

$$P(Etiquetas|Características) = \frac{P(Características|Etiquetas)*P(Etiquetas)}{P(Características)} \quad (14)$$

La metodología asociada al algoritmo de Naive Bayes es simple, para el caso en el que el algoritmo contenga sólo una característica, se describe a continuación:

1. Calcula la probabilidad asociada a las etiquetas de cada clase.
2. Obtener la probabilidad de los atributos para las clases designadas.
3. Aplicar la función probabilística de Bayes y calcular la probabilidad posterior.
4. Obtener la clase con la mayor probabilidad de ocurrencia.

2.7 Impulso adaptativo

El algoritmo de impulso adaptativo (Adaboost) fue propuesto en 1995 por Freund y Shapiro. El algoritmo recibe su nombre de dos palabras que lo describen fielmente, adaptativo (adaptive) e impulso (boost).

El algoritmo de impulso adaptativo es un modelo supervisado de inteligencia artificial que une distintos clasificadores simples con el propósito de generar un algoritmo robusto, el término generalmente usado para el algoritmo robusto es “fuerte” y para el algoritmo simple es “débil”.

El modelo que se describió anteriormente es un impulso conjunto, que pudiera tener ciertas similitudes con los modelos de árboles de decisión, sin embargo, este modelo opera secuencialmente, ya que tiende a aprender de los errores de los modelos débiles que anteceden en la secuencia, de esta forma aumenta el peso de los puntos de datos en los que se cometió una clasificación errónea, en la Figura 29 se ilustra [19].

En la Figura se muestran dos modelos débiles, en los que su entrada son un conjunto de datos, y su salida son el mismo conjunto de datos “clasificados”, que posteriormente, pasan al segundo modelo débil, pero el peso de este modelo cambia para reducir los errores cometidos por el modelo anterior.

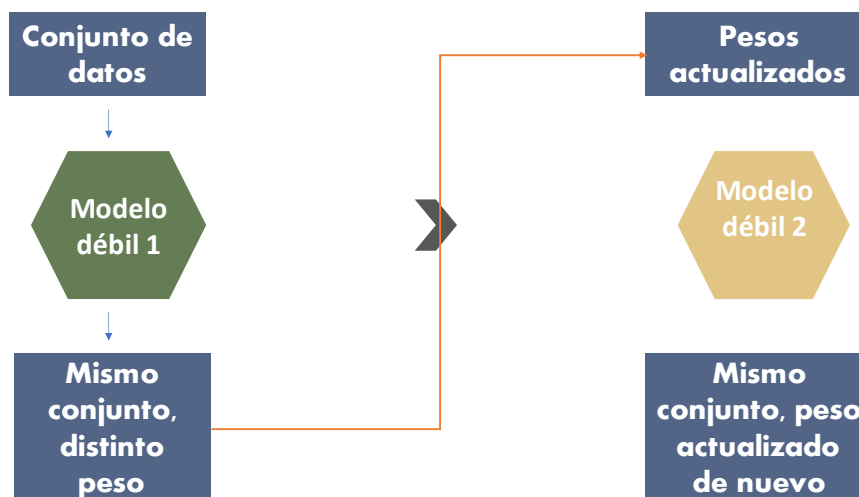


Figura 29: Secuencia del modelo, primer algoritmo “débil” (izquierda), segundo algoritmo “débil” (derecha).

2.7.1 Aprendizaje de impulso adaptativo

Como se mencionó en la sección anterior, el algoritmo de impulso adaptativo (Adaboost) aprende en base a los errores aumentando los pesos de los datos clasificados erróneamente. Impulso adaptativo es un modelo que trabaja con algoritmos “débiles”, estos algoritmos pueden ser bosque aleatorio, arboles de decisión, vecinos más cercanos, regresión lineal, entre otros más. El algoritmo predice en cada iteración, sumando los pesos, por lo que para este modelo se utilizará el algoritmo de árboles de decisión en conjunto con impulso adaptativo como ejemplo, a continuación, se describirá la serie de pasos a seguir para construir el modelo y su función matemática [19].

$$H(x) = \text{sen}(f(x)) = \text{sen}(\sum_{i=1}^B \alpha_b h_b(x)) \quad (15)$$

En donde:

B es la cantidad máxima de árboles.

b es cada arbol individual.

$h_b(x)$ es un clasificador debil.

α_b es el factor que combina el valor de los clasificadores debiles.

$\text{sen}(f(x))$ es la respuesta del clasificador.

$H(x)$ es el clasificador final.

La metodología para aplicar el algoritmo de aprendizaje supervisado de impulso adaptativo se describe a continuación:

1. Inicializar los pesos en los puntos de datos, el valor dado a cada peso es inversamente proporcional a la cantidad de pesos.
2. Entrenar el primer algoritmo “débil”.

3. Calcular el error ponderado del primer algoritmo “débil”, con lo que se obtiene la cantidad de predicciones realizadas incorrectamente.
4. Calcular el peso del segundo algoritmo “débil”, de acuerdo a la siguiente función:

$$W = T * \log\left(\frac{1-e}{e}\right) \quad (16)$$

En donde:

$W \rightarrow$ *Peso del algoritmo debil.*
 $T \rightarrow$ *Tasa de aprendizaje*
 $e \rightarrow$ *Error asociado a h_b*

5. Actualizar los pesos de los puntos de datos en los que se realizó una clasificación errónea.
6. Repetir del paso uno al paso cinco.

Los algoritmos como impulso adaptativo son técnicas de inteligencia artificial sumamente eficientes y efectivas, ya que permiten mejorar la precisión del algoritmo “débil”, sin embargo, existe el inconveniente de que los resultados son más complicados de interpretarse comparados con otros modelos, como lo es bosque aleatorio o árboles de decisión, precisamente es algo que caracteriza a los árboles de decisión, la facilidad de interpretar sus resultados.

2.8 Bosque aleatorio

Bosque aleatorio (Random Forest) es una técnica de inteligencia artificial propuesta por Kam Ho en 1995 y desarrollada por Breinman en 2001, el modelo comenzó como una combinación de dos algoritmos, los cuales son árbol de regresión y clasificación (Classification and regression tree) además, de remuestreo con reposición (Bootstrap aggregating). Con el objetivo de crear un algoritmo capaz de combinar árboles predictores como en el algoritmo de árboles de decisión, y obtener un valor conciso de vectores aleatorios con la misma distribución. El propósito de esto era entrenar múltiples clasificadores débiles de forma independiente, para finalmente obtener un resultado común en base a el resultado más frecuente, mejorando la precisión de los clasificadores individuales. El modelo se caracteriza por corregir el error de las predicciones, sin generar un único árbol, por el contrario, se construyen múltiples árboles, además, se establece un criterio de priorización de las variables con base a la predicción [19][23], este proceso se muestra en la Figura 30.

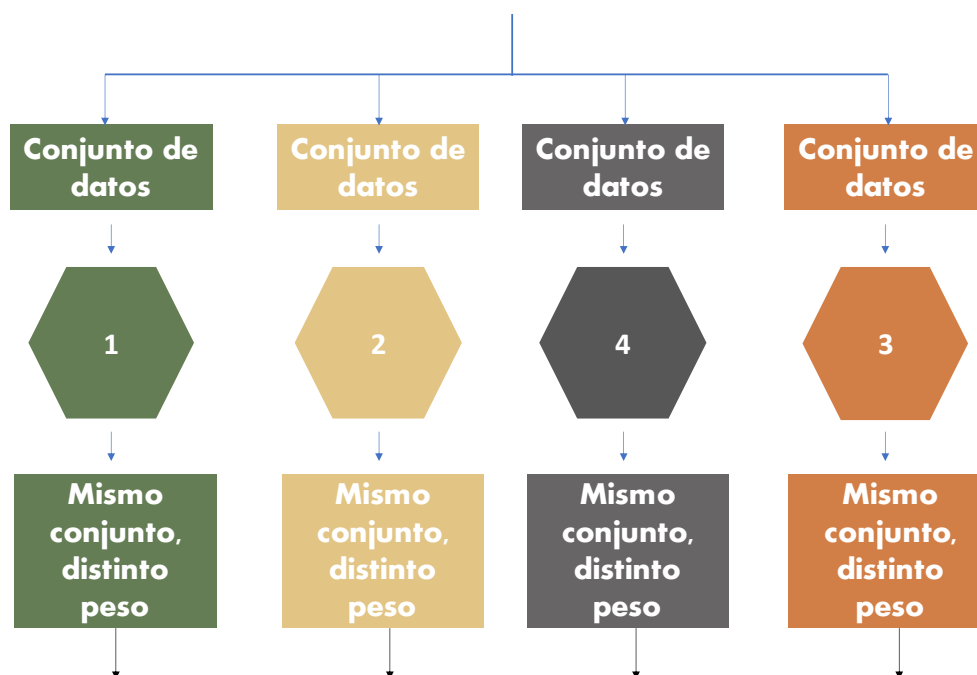


Figura 30: Secuencia del modelo de bosque aleatorio, con cada árbol de decisión.

2.8.1 Aprendizaje de bosque aleatorio

El algoritmo de Bosque aleatorio (Random Forest) es un modelo que crea múltiples árboles de decisión con base en un conjunto de datos, con el objetivo de obtener un resultado conjunto, de los datos obtenidos de cada árbol de decisión por separado. Este algoritmo se desarrolla de forma general en base a la siguiente metodología, sin embargo, esta podría cambiar según la problemática planteada [19][23], en la Figura 31.

1. Extraer una muestra del tamaño que se requiera, de los datos de entrenamiento, estos deben de ser aleatorios y serán utilizados para calcular la cantidad datos clasificados erróneamente.
2. Construir un árbol de decisión con los datos que no se tomaron de la muestra, seleccionando las variables de forma aleatoria, dividiendo el nodo en dos nodos subsecuentes, creciendo hasta su máxima extensión.
3. Calcular el error de clasificación del algoritmo en base a las observaciones no estimadas en cada iteración, promediando lo resultados obtenidos.

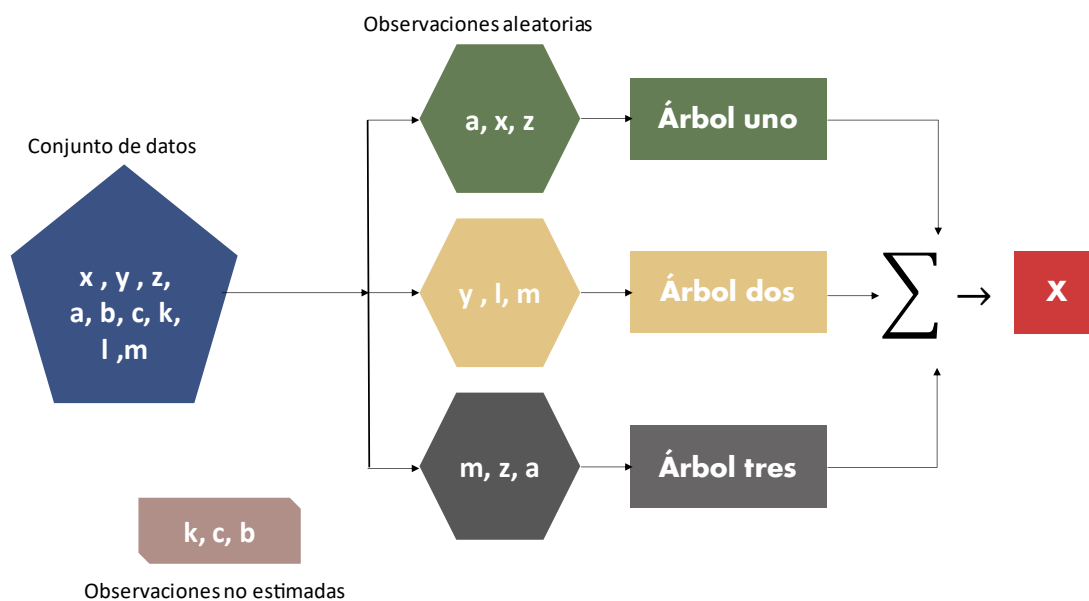


Figura 31: Proceso de aprendizaje del algoritmo de bosque aleatorio.

2.9 Algoritmos de inteligencia artificial con WEKA®

El Machine Learning es una rama de la inteligencia artificial, en la que se utilizan algoritmos y distintas técnicas, que tienen como objetivo la obtención de conocimiento en base a conjuntos de datos para su análisis y visualización. Existen diversos softwares de apoyo para el aprendizaje automático que aplican técnicas de inteligencia artificial para analizar y modelar datos. WEKA® (Waikato Environment for Knowledge Analysis) es un software de código abierto, que presenta un entorno para el análisis del conocimiento desarrollado en la Universidad de Waikato en Nueva Zelanda en 1999. Es ampliamente utilizado en minería de datos y aprendizaje automatizado, aplicado en tareas de regresión, clasificación y de agrupación. WEKA® permite cargar y ejecutar conjuntos de datos en un formato propio, ARFF (Attribute-Relation File Format) [20]. Es un software sumamente versátil y eficiente ya que arroja resultados sólidos en un corto periodo de tiempo.

WEKA® contiene cuatro aplicaciones en su interfaz gráfica, esta se puede observar en la Figura 32, las cuales son:

- Simple CLI (Interfaz de línea de comandos): Permite al usuario introducir comandos, para la ejecución de tareas.
- Workbench (Banco de trabajo): Permite al usuario visualizar de forma completa el panel de control de todas las aplicaciones de WEKA®.
- Knowledge flow (Flujo de conocimiento): Permite al usuario utilizar un panel de control para visualizar gráficamente tareas de clasificación, preprocesado de datos, etc.
- Experimenter (Experimentador): Permite al usuario obtener resultados de índices estadísticos de un gran conjunto de datos.
- Explorer (Explorador): Permite al usuario operar sobre un conjunto de datos, permitiendo realizar tareas tales como la clasificación, preprocesado de datos, asociaciones, visualización de datos, etc.

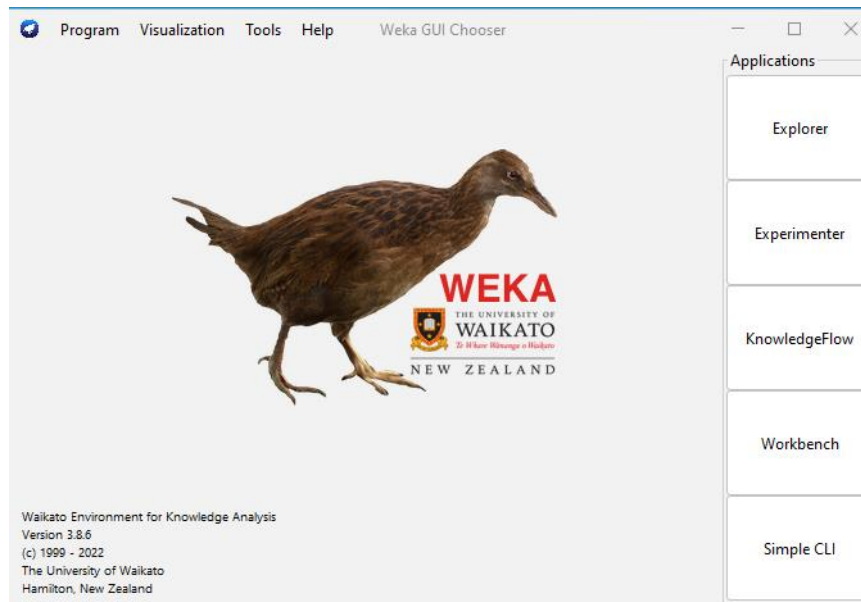


Figura 32: Interfaz gráfica de usuario (GUI) del software WEKA®.

Capítulo 3

3 Metodología.

3.1 Metodología propuesta

En el presente trabajo de tesis se ha hecho mención de forma general, sobre la metodología del procesamiento digital de imágenes y de la clasificación de estas mediante algoritmos de aprendizaje supervisado. En este capítulo se plantea la metodología específica para el presente caso, en el que se evalúa la viabilidad celular mediante modelos de inteligencia artificial, además, se muestra la experimentación que se llevó a cabo a lo largo de este trabajo de tesis. La metodología propuesta para la presente investigación contiene 5 etapas, las cuales pueden ser apreciadas en la Figura 33. Las etapas son las siguientes:

- I. Preprocesado de datos.
- II. Detección de objetos.
- III. Extracción de características.
- IV. Entrenamiento.
- V. Clasificación.

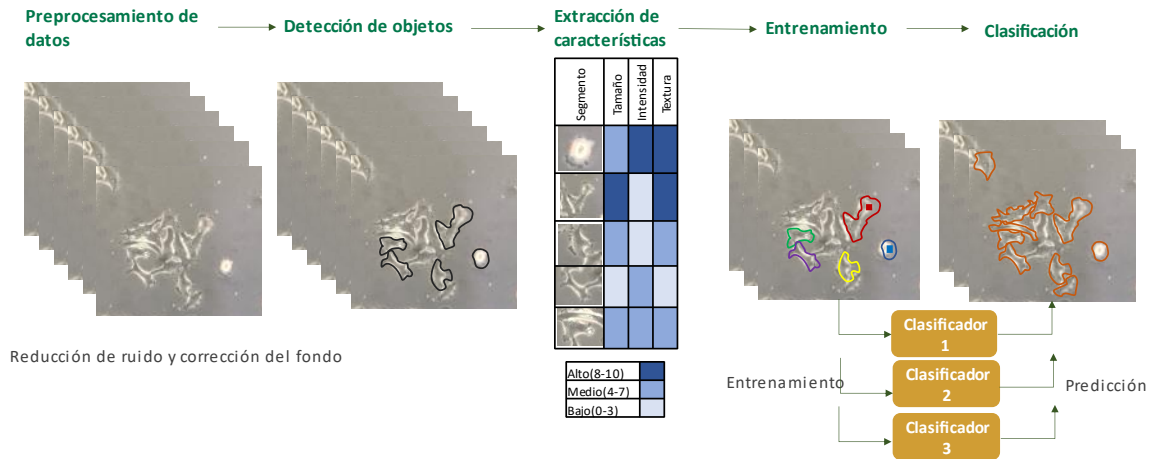


Figura 33: Metodología propuesta, la cual consta de cinco etapas, el preprocesamiento de datos, la detección de objetos, la extracción de características, el entrenamiento y la clasificación.

Para el procesamiento digital de imágenes, la detección de objetos, la extracción de características, el entrenamiento de los algoritmos y la clasificación de imágenes, se hizo uso de *FIJI®* en versión *Java 8*. Para ello, en dicho software se utilizaron los menús *procesar* (process), *analizar* (analyze) y *complementos* (plugins). Para ligar *FIJI®* y *WEKA®* se implementó el complemento Segmentación Entrenable de Weka (TWS). En lo referente a la evaluación de los algoritmos de inteligencia artificial se hizo uso de *WEKA®* en versión 3.8.6, para lo cual se utilizó la aplicación de *explorador* (explorer), de trabajar con esta sección se obtuvieron las matrices de confusión. Dichas matrices se utilizaron para evaluar el desempeño de los algoritmos y realizar el comparativo.

Las etapas mencionadas anteriormente fueron realizadas en un ordenador portátil Dell® Precision M6800 con procesador Intel® Core i7 de cuarta generación a 2.70 GHz con 16 GB de memoria RAM y un disco duro de estado sólido de 128 GB, en un sistema operativo Microsoft® Windows 10.

3.2 Preprocesamiento de imágenes

En el presente trabajo de tesis se generó una base de datos de imágenes con células de cáncer de mama de la línea celular MDA-MB-231 obtenidas en ambiente de laboratorio tomadas al microscopio.

Cada imagen de la base de datos original fue fragmentada para generar una base de datos más amplia y con imágenes de tamaño manejable por las limitaciones en los recursos de cómputo disponibles para el desarrollo del presente trabajo de tesis. Así, se obtuvieron alrededor de 42,000 imágenes de 126 x 126 píxeles. A este conjunto de datos se le llamará Base de datos de referencia (BDR) a lo largo de este documento.

Las imágenes que se encuentran en esta base de datos, serán utilizadas en la primera etapa, la cual consiste en preprocesar imágenes de cultivos celulares, con el objetivo de reducir el ruido en la imagen, corregir el fondo, el enfoque y el contraste, lo que puede ser generado por factores que se presentan durante la toma de la fotografía, con el propósito de brindar imágenes con las características necesarias para la siguiente etapa [21].

Para ello, se seleccionan las regiones de interés en las imágenes de la BDR (Células de cáncer de mama) y se aplican los filtros espaciales: Difusión anisotrópica, Agudizado, y Kuwahara, en cada imagen. En la Figura 34, se muestra el efecto de la aplicación de los filtros espaciales arriba mencionados en células de cáncer de mama. Las ilustraciones paso a paso de cómo se realiza este procedimiento, se muestra en el apéndice.

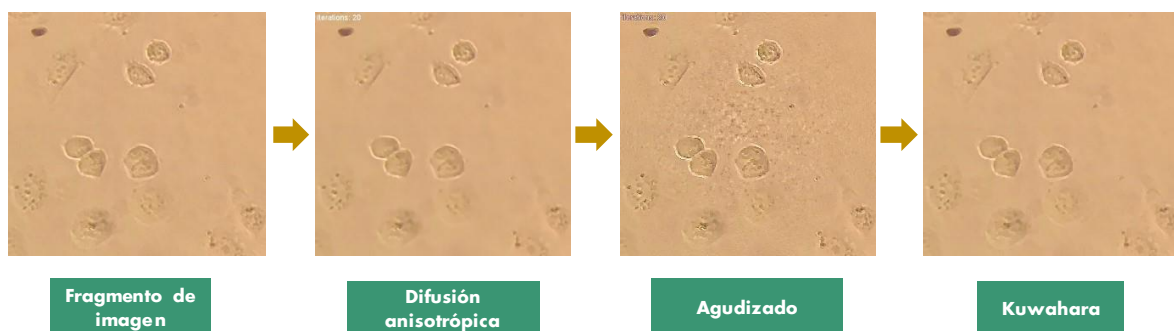


Figura 34: Metodología implementada en el preprocesamiento digital de imágenes celulares, para la línea celular MDA-MB-231.

3.3 Detección de objetos

Con el objetivo de detectar los objetos de interés en la imagen, se continúa con la segunda etapa de la metodología propuesta. La presente etapa consiste en detectar las células y separarlas del fondo. Esto se realiza mediante operadores gradientes y de morfología matemática, con el propósito de generar enmascarados que permitan diferenciar los objetos de interés de cada imagen, a fin de extraer las características en la tercera etapa de la metodología. En la Figura 35, se puede apreciar el resultado de la detección de objetos por medio del operador de Sobel. En la Figura 36 se puede apreciar la morfología matemática aplicada con el dilatador y el relleno de objetos, para posteriormente convertirlo en una máscara [5] [21]. Cabe mencionar que los resultados de esta operación fueron evaluados por un experto en el reconocimiento de células viables de cáncer de mama y con experiencia en el trabajo de esta línea celular. Las ilustraciones paso a paso de cómo se realiza este procedimiento, se muestra en el apéndice.

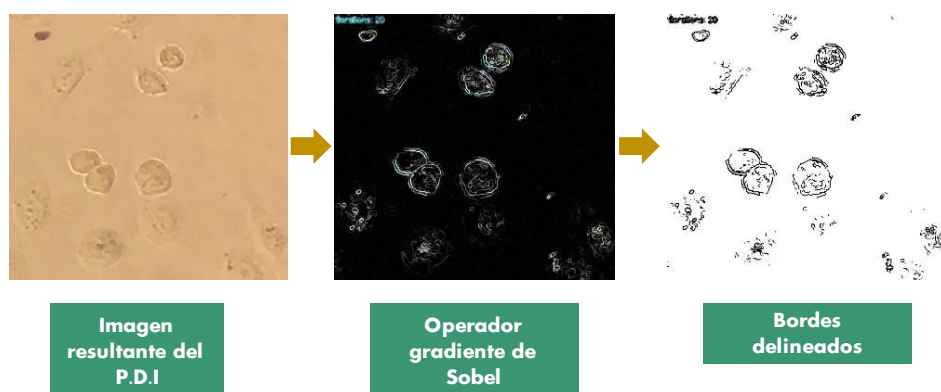


Figura 35: Metodología implementada en la detección de bordes.



Figura 36: Metodología empleada en la realización del enmascarado.

3.4 Extracción de características

En la tercera etapa de la presente metodología, se extraen diversas características morfológicas de la imagen, que fueron definidas en el capítulo 2. Esta clase de características le brindan al usuario información acerca de las partículas (debris) que se hayan detectado como objetos en la etapa anterior, utilizando los píxeles como base. El objetivo de extraer información acerca de las características es el de parametrizar las imágenes de la base de datos generada, para categorizarlas y etiquetarlas, esto les permitirá a los algoritmos de inteligencia artificial ser entrenados adecuadamente.

Esto último compete a la cuarta etapa de la metodología propuesta [5] [21]. Las características morfológicas que serán utilizadas pueden apreciarse en la Figura 37, y son mencionadas en la siguiente Tabla 3:

Tabla 3: Características morfológicas de las células de cáncer de mama y su descripción

Características	Descripción
Perímetro	Brinda al usuario la longitud del límite exterior de la imagen o de la sección que se designó.
Fracción de área	Indica al usuario el porcentaje de píxeles que se han resaltado en la sección designada o en la imagen completa.
Ajustar elipse	Le permite al usuario ajustar la elipse que se marca con el parámetro centroide, utilizando los ejes primarios y secundarios de la elipse, y el ángulo entre el eje principal y la línea paralela al eje "x".
Descriptor de Figura	Le brinda al usuario información acerca de los parámetros circularidad, solidez, relación de aspecto y redondez.
Densidad integrada	Muestra al usuario el producto del área y el valor de la media en escala de grises y la suma de los valores de los píxeles en la sección designada o de la imagen completa.
Feret	También conocido como calibre máximo, es la máxima distancia entre dos puntos a lo largo del límite de una sección. Brinda al usuario información acerca de los parámetros Feret, ángulo de Feret, mínimo de Feret, y las coordenadas X y Y de Feret.
Área	Permite al usuario calibrar la imagen espacialmente en píxeles o en milímetros.

Las características mencionadas anteriormente brindan información acerca de la forma y tamaño de las células, la cuantificación de cada una de las características permite discriminar las células viables de las no viables, además, cada uno de los parámetros será utilizado como la entrada del clasificador en la cuarta etapa de la metodología, que se refiere al entrenamiento. Los pasos a seguir para la realización de la tercera etapa, son los siguientes:

- ConFigurar las características que serán extraídas, mediante la ventanilla de “establecer medidas”.
- Extraer las características de la imagen, mediante la ventanilla de “analizar partículas”.

En la Figura 38, se pueden apreciar los resultados visuales del enmascarado de la extracción.

Las ilustraciones paso a paso de cómo se realiza este procedimiento, se muestran en el apéndice.

	Label	Area	X	Y	Perim.	Major	Minor	Angle	Circ.	Feret	IntDen	Median	%Area	RawIntDen	Slice	FeretX	FeretY	FeretAngle	MinFeret	AR	Round	Solidity
1	IMG_9415_2_3-Iter20	10453	328.345	122.248	468.250	154.592	86.092	50.172	0.599	155.795	2665515	255	100	2665515	1	277	181	48.122	98.302	1.796	0.557	0.889
2	IMG_9415_2_3-Iter20	5605	130.988	140.128	455.765	96.978	73.589	47.797	0.339	117.597	1429275	255	100	1429275	1	88	176	33.555	91.103	1.318	0.759	0.777
3	IMG_9415_2_3-Iter20	836	404.142	221.577	111.941	39.552	26.912	26.929	0.838	43.600	213180	255	100	213180	1	387	235	36.607	28.000	1.470	0.680	0.968
4	IMG_9415_2_3-Iter20	10079	181.284	338.310	414.635	133.192	96.349	135.577	0.737	136.620	2570145	255	100	2570145	1	130	293	132.330	99.017	1.382	0.723	0.934
5	IMG_9415_2_3-Iter20	8171	312.465	353.762	361.037	104.175	99.867	76.142	0.788	111.396	2083605	255	100	2083605	1	274	397	49.733	101.000	1.043	0.959	0.949
6	IMG_9415_2_3-Iter20	749	464.051	391.360	107.113	36.457	26.159	39.914	0.820	40.460	190995	255	100	190995	1	449	405	39.987	28.988	1.394	0.718	0.944
7	IMG_9415_2_3-Iter20	493	296.015	419.557	84.142	25.658	24.465	80.691	0.875	30.414	125715	255	100	125715	1	285	430	46.332	22.000	1.049	0.953	0.989
8	IMG_9415_2_3-Iter20	298	551.500	421.000	65.314	20.369	18.628	113.578	0.878	24.759	75990	255	100	75990	1	543	412	133.363	17.000	1.093	0.915	0.990
9	IMG_9415_2_3-Iter20	284	524.947	426.049	64.485	20.248	17.859	98.060	0.858	24.083	72420	255	100	72420	1	517	417	131.634	16.000	1.134	0.882	0.993
10	IMG_9415_2_3-Iter20	2770	333.829	452.140	232.912	67.939	51.912	20.189	0.642	73.539	706350	255	100	706350	1	303	466	22.380	57.046	1.309	0.764	0.875
11	IMG_9415_2_3-Iter20	5763	197.962	507.397	604.794	100.836	72.768	21.286	0.198	123.458	1469565	255	100	1469565	1	139	532	35.106	86.512	1.386	0.722	0.752

Figura 37: Características morfológicas extraídas mediante el software FIJI®.

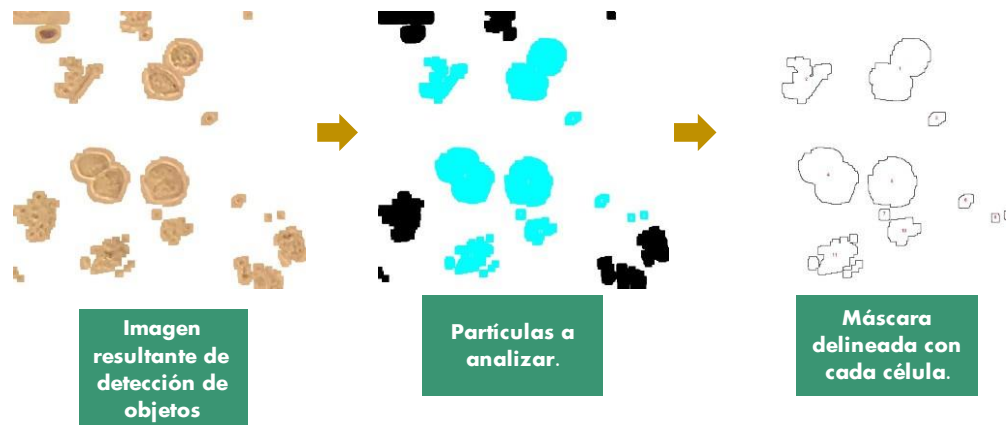


Figura 38: Metodología empleada en la extracción de características morfológicas de las células.

3.5 Entrenamiento

Con el objetivo de identificar automáticamente las células viables en las imágenes, se hará uso de algoritmos de inteligencia artificial. Para obtener una clasificación con un alto porcentaje de instancias clasificadas correctamente, el algoritmo en cuestión debe ser entrenado con un conjunto de datos de entrenamiento, que, en el presente caso, son imágenes, que han sido etiquetadas previamente por un experto en el reconocimiento de células viables de cáncer de mama y con experiencia en el trabajo de esta línea celular. Para estas imágenes se definieron las siguientes etiquetas:

- Célula viable.
- Célula no viable.
- Fondo de caja de Petri.
- Debris.
- Borde de la lente.
- Fondo negro.

Los tres algoritmos de inteligencia artificial bajo comparación se entrenaron con los vectores de características extraídos del conjunto de datos de entrenamiento generados con WEKA® y almacenados en los archivos con formato “. arff”. El proceso de entrenamiento se lleva a cabo en el complemento de Segmentación Entrenable de Weka (TWS).

Los pasos que se siguieron para realizar el entrenamiento del algoritmo se aprecian en la Figura 39, y se muestran a continuación:



Figura 39: Metodología empleada en el entrenamiento del algoritmo en TWS.

Las ilustraciones paso a paso de cómo se realiza este procedimiento, se muestran en el apéndice.

3.6 Clasificación

La quinta y última etapa de la metodología propuesta consiste en discriminar las células viables de las que no lo son, en función del entrenamiento del algoritmo y del etiquetado, en conjunto con las características morfológicas extraídas. La clasificación será llevada a cabo por tres algoritmos de inteligencia artificial, estos son:

- Naive Bayes.
- Impulso adaptativo.
- Bosque aleatorio.

La elección de los clasificadores fue realizada con base a su buen desempeño reportado en la literatura especializada [5] [20].

Para el proceso de clasificación se particionó la base de datos, tomando el 80% de las imágenes para para entrenamiento y el 20% restante para prueba. La predicción arrojada en esta área de prueba por parte del algoritmo, se compara con la información recabada mediante la inspección directa de las imágenes por parte de un experto en el reconocimiento de células viables de cáncer de mama y con experiencia en el trabajo de esta línea celular. Al final, se evalúa si es correcta o no la predicción comparándola con la opinión del experto para establecer el resultado. Los pasos a seguir para realizar la clasificación de imágenes celulares son los siguientes:

- Cargar el modelo supervisado de inteligencia artificial, el archivo “. model”.
- Cargar los vectores de datos con los atributos e instancias, el archivo “. arff”.
- Aplicar el clasificador en las imágenes.

Las ilustraciones paso a paso de cómo se realiza este procedimiento, se muestra en el apéndice.

En la Figura 40 se muestra el resultado del proceso de etiquetado y la clasificación de la imagen mediante el mapa de probabilidad, además de la imagen segmentada. Mientras que en la Figura 41, se aprecian las clases asignadas a cada etiqueta y la identificación de cada clase sobre la imagen.

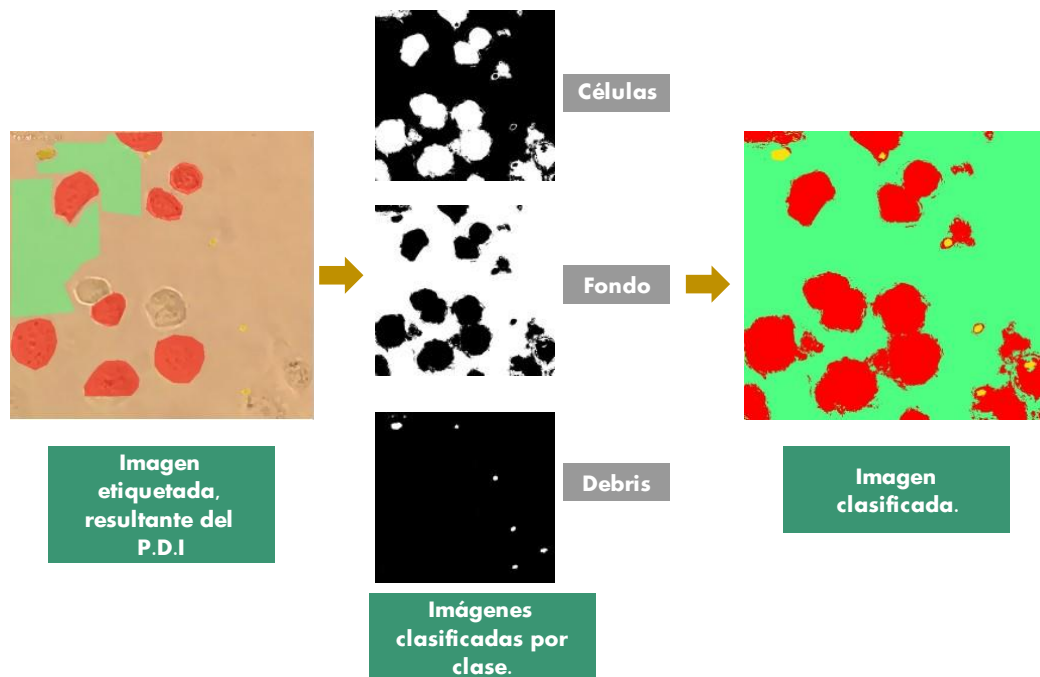


Figura 40: Imágenes etiquetadas y clasificadas mediante mapa de probabilidades y segmentación

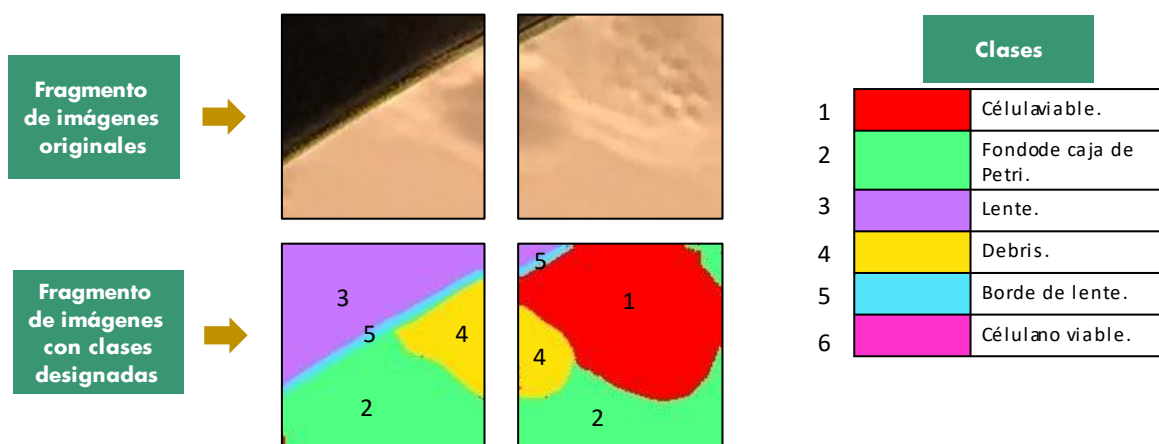


Figura 41: Imágenes fragmentadas e imágenes clasificadas con cada clase designada.

3.7 Evaluación de los clasificadores

Con el propósito de obtener un parámetro que nos indique el desempeño de los clasificadores se optó por evaluar a los clasificadores de dos formas, la primera es de forma cuantitativa, en la que se tomará como referencia la matriz de confusión resultante de la aplicación de cada uno de los algoritmos. La segunda forma es la cualitativa, en la que las imágenes serán la referencia para su evaluación.

La matriz de confusión nos brinda información acerca de la precisión y sensibilidad que el algoritmo presenta, además nos indica la cantidad de instancias que el modelo clasificó correctamente y cuantas otras se confundieron con el resto de clases.

La matriz de confusión se define como una herramienta visual que permite al usuario analizar concretamente el desempeño de un algoritmo o modelo de aprendizaje supervisado. La matriz de confusión contiene 4 parámetros para indicar de qué manera se clasificaron las instancias, estas son verdadero negativo (TN), falso positivo (FP), falso negativo (FN) y verdadero positivo (TP) [24].

La matriz debe interpretarse de la siguiente forma:

- Cada renglón y columna contiene una clase definida.
- La diagonal formada, corresponde a los valores que el algoritmo estimó de forma correcta.
- Cualquier valor encontrado fuera de la diagonal principal, se consideran valores erróneos, en los que el modelo confundió las instancias.

Matriz de confusión						
a	b	c	d	e	f	Clasificado como:
10	0	0	1	0	2	a = Célula viable
0	20	0	0	0	0	b = Fondo
0	0	50	0	0	0	c = Lente negro
0	0	0	15	0	0	d = Debris
0	0	0	0	18	0	e = Borde lente
5	0	0	0	0	20	f = Célula no viable

Figura 42: Ejemplo de una matriz de confusión de 6x6.

En la Figura 42, se muestra un ejemplo de una matriz de confusión. En el ejemplo que se ilustra, la matriz de confusión contiene 6 clases, de las cuales Fondo, Lente negro y Borde lente no presentan ninguna instancia clasificada erróneamente, sin embargo, en las clases Célula viable, Debris y Célula no viable, si existen clasificaciones erróneas. En la primera columna, en el primer y sexto renglón se encuentran los valores 10 y 5, el valor de 10 corresponde a las instancias clasificadas correctamente como Célula viable, sin embargo, el valor de 5, corresponde a las instancias clasificadas incorrectamente, con las que existe, en este ejemplo, la confusión por parte del algoritmo, entre Célula viable y Célula no viable.

En el presente trabajo de tesis, la evaluación de los clasificadores es llevada a cabo en el software de WEKA®, en la sección de “clasificar” (Classify). Los pasos que se siguieron para ello son los siguientes:

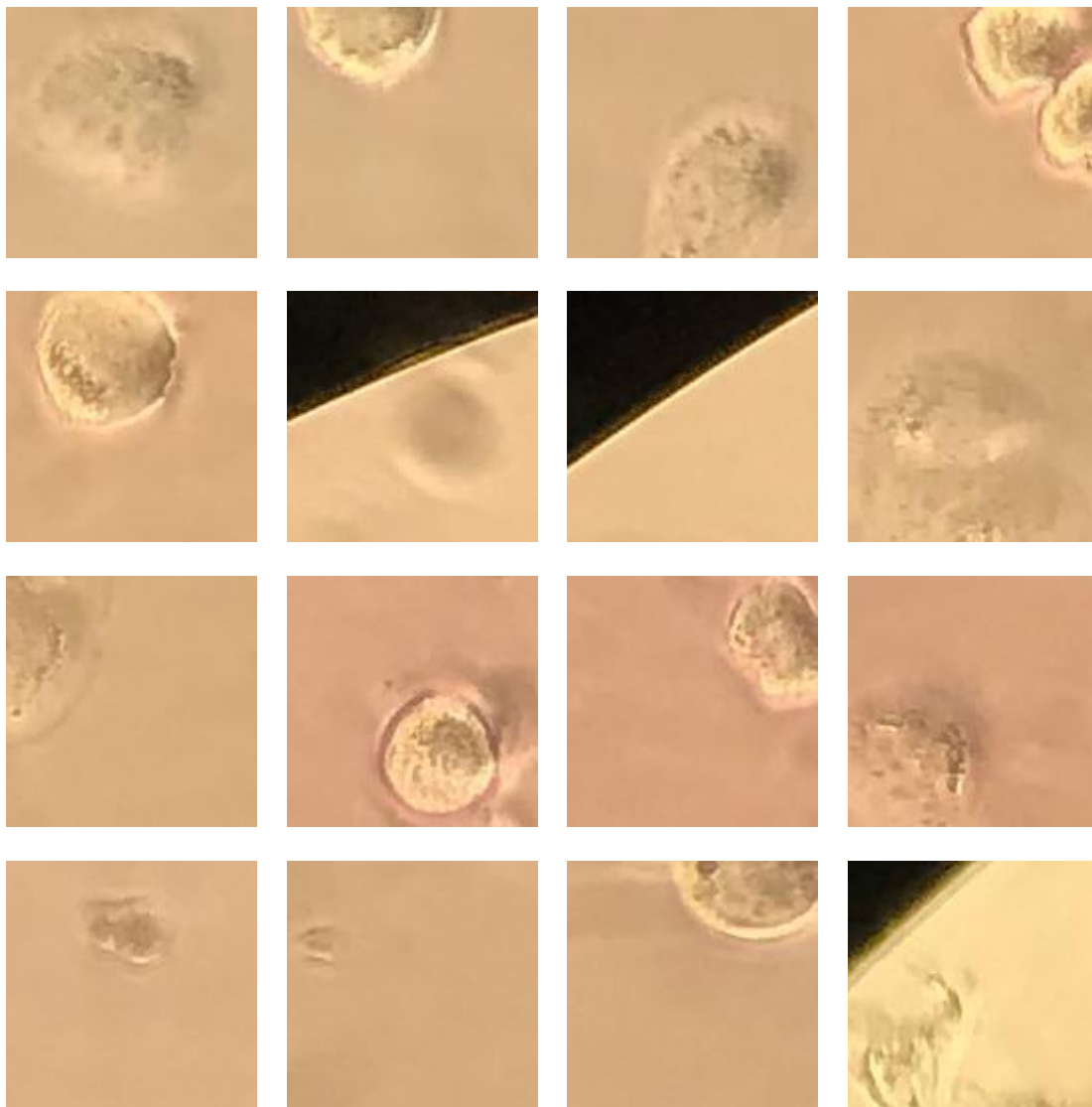
- Inicializar WEKA®.
- Inicializar la aplicación de “explorador” (explorer) del menú de WEKA®.
- Seleccionar la pestaña de “preprocesar” (preprocess).
- Abrir el conjunto de datos, con extensión “. arff”.
- Seleccionar la pestaña de “clasificar”.
- Seleccionar el modelo de aprendizaje supervisado.
- Indicar la partición de datos.
- Obtener la matriz de confusión.

Capítulo 4

4 Resultados

4.1 Base de datos generada

En el presente trabajo de tesis se han utilizado imágenes de la línea celular MDA-MB-231, tomadas de un microscopio invertido modelo Vert A1, marca Axio, con lo que construyó la base datos, la cual contiene alrededor de 42,000 imágenes de la línea celular mencionada anteriormente. Dado el gran número de imágenes en la base de datos y con el propósito de ilustrar paso a paso los resultados obtenidos en cada parte de la metodología utilizada en este proyecto de tesis, se seleccionó un conjunto de 20 imágenes, las cuales se muestra el conjunto de imágenes en la Figura 43:



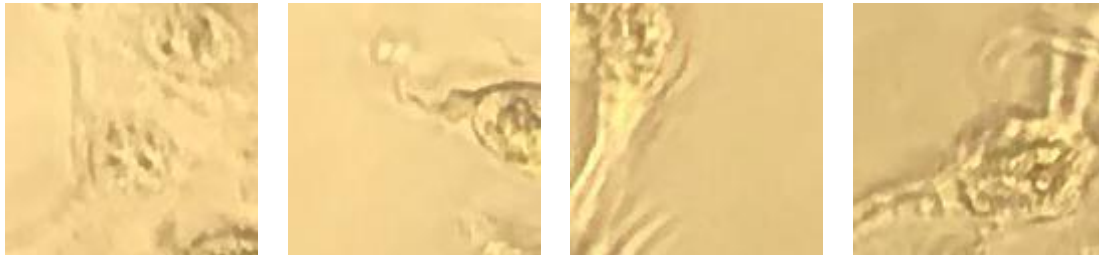


Figura 43: Conjunto de imágenes originales de la base de datos.

4.2 Resultados obtenidos por procesamiento digital de imágenes

Los resultados obtenidos de la primera y segunda etapa de la metodología propuesta, que comprenden el preprocesamiento digital de imágenes y la detección de objetos, se mostrarán a continuación. En la Figura 44, se muestra el resultado de aplicar el filtro de difusión anisotrópica a 20 imágenes de la base de datos de la línea celular MDA-MB-231, usadas para ilustrar los resultados, en ellas se puede observar, que el fondo de la imagen se vuelve más uniforme, con una menor cantidad de ruido, conservando los bordes de las células.

En la Figura 45, se puede observar el resultado de aplicar en las imágenes el filtro de agudizado se vuelve visible en los contornos de las células y en los detalles de separación entre el fondo de caja de Petri y la célula, mientras que en la Figura 46, el resultado de aplicar el filtro Kuwahara, las células se acentúan aún más, sus bordes se vuelven más definidos y el fondo se aprecia con más uniformidad, lo que sienta las bases para que la siguiente etapa realice se efectuó correctamente. Se puede comentar que, comparado con las imágenes originales, el resultado de este proceso muestra que se logra eliminar el ruido en la imagen casi en su totalidad, además, los bordes se vuelven claramente visibles, encontrando una clara diferencia entre el fondo, las células y los restos celulares.

Ahora, en la Figura 47 se ilustra el efecto de aplicar el operador gradiente de Sobel en las imágenes. En ellas se puede observar que los bordes de las células y los restos celulares son definidos, marcando los límites entre los objetos de interés que son las células y el fondo. Lo que le permitirá a la siguiente etapa de la metodología, extraer las características morfológicas.

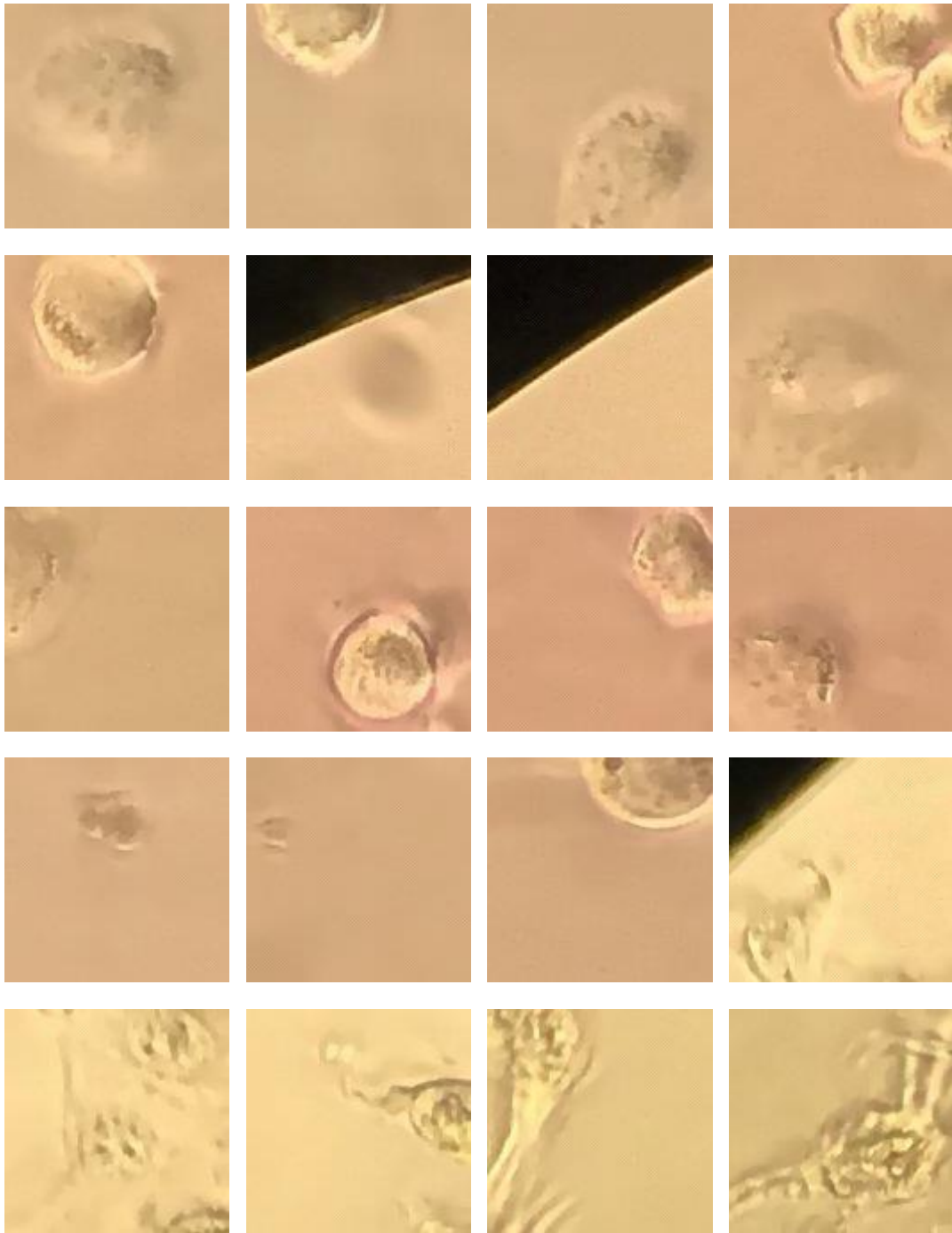


Figura 44: Imágenes con filtro de difusión anisotrópica.

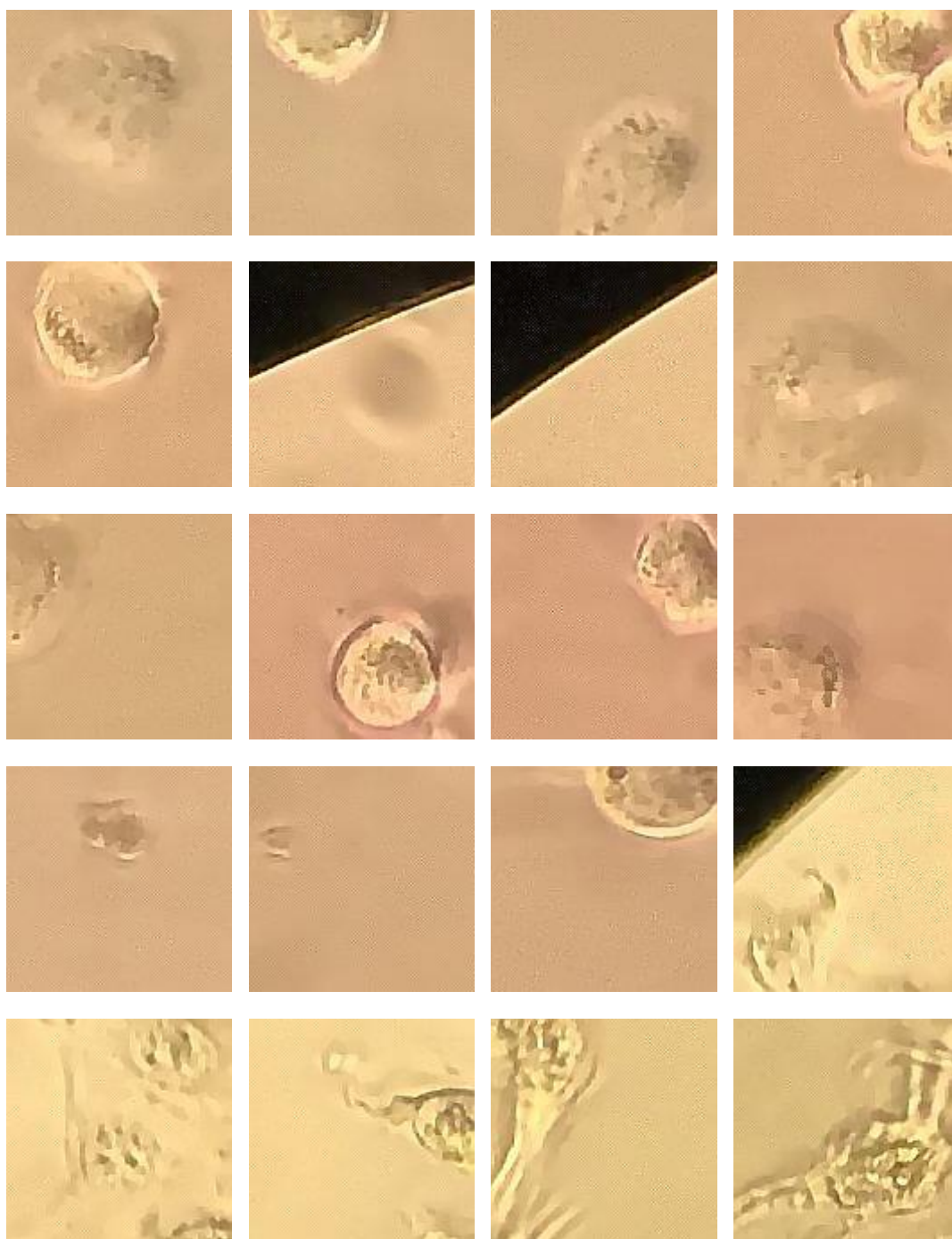


Figura 45: Imágenes con filtro de agudizado.

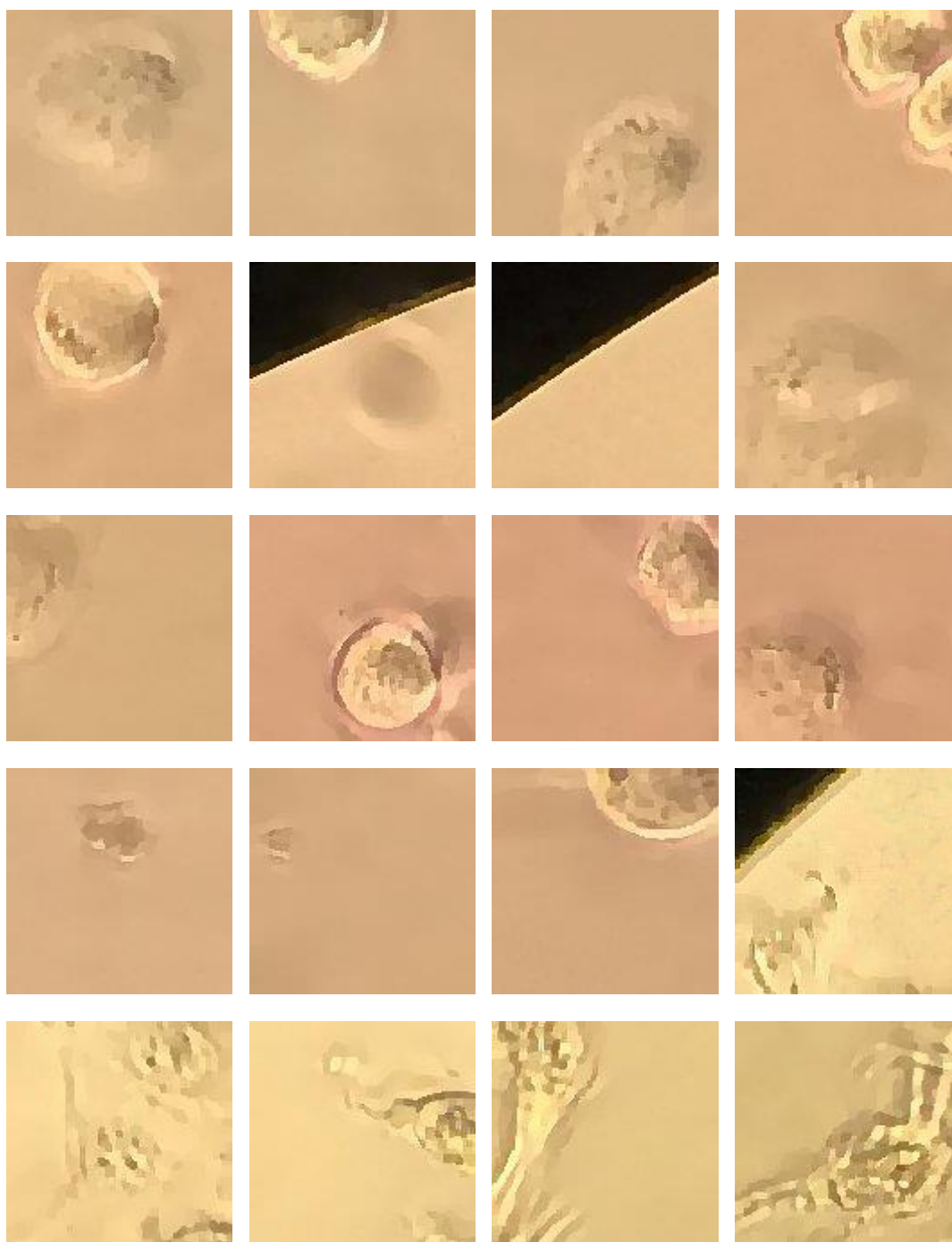


Figura 46: Conjunto de imágenes con filtro Kuwahara.

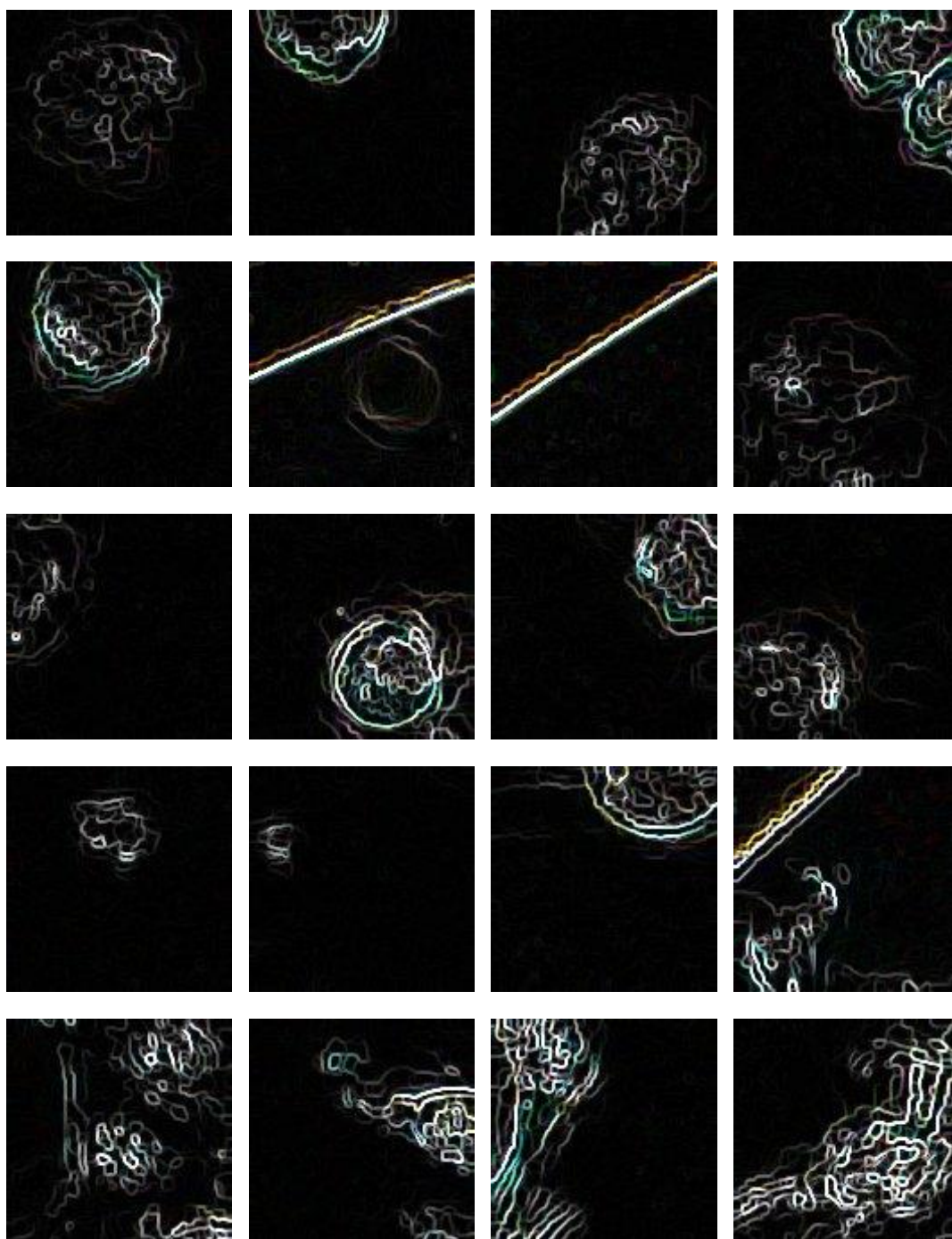


Figura 47: Imágenes con filtro Sobel.

4.3 Resultados obtenidos al aplicar la extracción de características

Tabla 4: Tabla de las características extraídas del conjunto de imágenes.

	Label	Area	X	Y	Perim.	Major	Minor	Angle	Circ.	Feret	IntDen	Median	%Area	RawIntDen	Slice	FeretX	FeretY	FeretAngle	MinFeret	AR	Round	Solidity	MinThr	MaxThr
1	img24_11_8-i	8333	60.576	50.154	504	111.827	94.878	42.739	0.412	178.191	2124915	255	52.488	2124915	1	0	0	135	126	1.179	0.848	1	255	255
2	img24_11_9-i	3587	44.656	21.194	248.651	89.928	50.786	0.662	0.729	93.005	914685	255	98.193	914685	1	1	0	163.124	50.5	1.771	0.565	0.95	255	255
3	img24_11_10-i	6851	79.971	85.118	339.695	106.077	82.233	39.003	0.746	122.745	1747005	255	98.96	1747005	1	31	125.5	43.184	85.997	1.29	0.775	0.959	255	255
4	img30_9_14-i	5440	93.848	39.231	344.468	103.097	67.184	120.18	0.576	124.201	1387200	255	99.817	1387200	1	50.333	2	127.147	80.333	1.535	0.652	0.904	255	255
5	img30_11_15	7096	51.923	39.665	364.583	102.032	88.55	11.429	0.671	111.057	1809480	255	99.495	1809480	1	8.333	1	147.502	88.083	1.152	0.868	0.933	255	255
6	img31_1_12-i	1243	66.849	36.619	293.031	143.192	11.053	22.809	0.182	135.373	316965	255	93.529	316965	1	1.667	69	25.682	13.636	12.96	0.077	0.768	255	255
7	img31_2_10-i	1501	62.632	45.629	362.617	160.714	11.892	33.621	0.143	153.835	382755	255	91.357	382755	1	0.333	91.333	35.653	16.111	13.52	0.074	0.639	255	255
8	img31_8_9-it	8913	56.501	79.988	415.895	115.57	98.195	151.38	0.648	124.069	2272815	255	99.21	2272815	1	16.333	39.333	137.177	99.243	1.177	0.85	0.925	255	255
9	img31_10_10	4158	24.365	40.551	278.147	94.532	56.004	77.589	0.675	100.817	1060290	255	99.355	1060290	1	0	89	54.182	59.643	1.688	0.592	0.945	255	255
10	img32_8_16-i	5301	78.242	88.377	280.009	82.63	81.682	123.75	0.85	93.497	1351755	255	98.715	1351755	1	53	52	128.341	78.225	1.012	0.989	0.962	255	255
11	img32_9_16-i	3019	102.72	33.804	226.238	74.046	51.912	99.798	0.741	80.035	769845	255	96.98	769845	1	82.667	0.667	121.373	53.152	1.426	0.701	0.928	255	255
12	img32_10_12	5521	40.922	87.054	324.375	88.458	79.468	165.28	0.659	103.317	1407855	255	98.996	1407855	1	-0.25	58.5	152.629	77.726	1.113	0.898	0.911	255	255
13	img42_7_13-i	1937	61.35	34.894	191.189	51.757	47.65	134.39	0.666	58.88	493935	255	93.665	493935	1	36.333	17.667	128.103	52.444	1.086	0.921	0.846	255	255
14	img42_10_11	569	15.793	42.328	95.325	29.28	24.743	91.263	0.787	32.351	145095	255	94.362	145095	1	8.25	55.75	53.484	25.461	1.183	0.845	0.905	255	255
15	img42_12_10	4233	84.97	25.536	260.798	93.23	57.81	170.92	0.782	99.56	1079415	255	98.442	1079415	1	38.333	0	150.517	55.667	1.613	0.62	0.976	255	255
16	img85_2_9-it	716	104.23	23.207	211.323	39.906	22.845	94.841	0.201	59.053	182580	255	66.481	182580	1	96	1.5	92.426	31.84	1.747	0.572	0.478	255	255
16_2	img85_2_9-it	794	23.913	20.739	155.484	63.406	15.944	39.53	0.413	71.171	202470	255	90.125	202470	1	3	49	41.868	17.895	3.977	0.251	0.848	255	255
17	img85_5_15-i	7520	59.588	47.503	423.987	121.283	78.946	154.53	0.526	125.143	1917600	255	99.13	1917600	1	18.5	2.5	127.587	84.79	1.536	0.651	0.871	255	255
18	img85_6_14-i	7131	57.245	52.96	357.604	118.962	76.322	97.811	0.701	125.345	1818405	255	97.525	1818405	1	27	0.5	119.121	78.535	1.559	0.642	0.92	255	255
19	img85_12_15	3620	22.841	67.45	290.882	89.98	51.224	76.829	0.538	106.354	923100	255	98.05	923100	1	-0.667	126.67	69.066	54.944	1.757	0.569	0.876	255	255
20	img197_7_16	4105	35.647	71.99	292.586	81.153	64.405	154.7	0.603	88.408	1046775	255	95.956	1046775	1	7	33	127.648	71.289	1.26	0.794	0.854	255	255

En los apartados anteriores, se hizo mención del conjunto de imágenes tomadas de la base de datos para ilustrar los resultados de la aplicación de las técnicas de procesamiento digital de imágenes, con base en la metodología propuesta. En esta sección, se ilustran los resultados obtenidos de la extracción de características del mismo conjunto de imágenes, en la Tabla 4 se muestran los resultados expresados numéricamente y en la Figura 48. En la Tabla se puede observar las características morfológicas extraídas, mientras que, en la Figura se aprecian las secciones de la imagen que fueron analizadas y tomadas para la extracción de características.

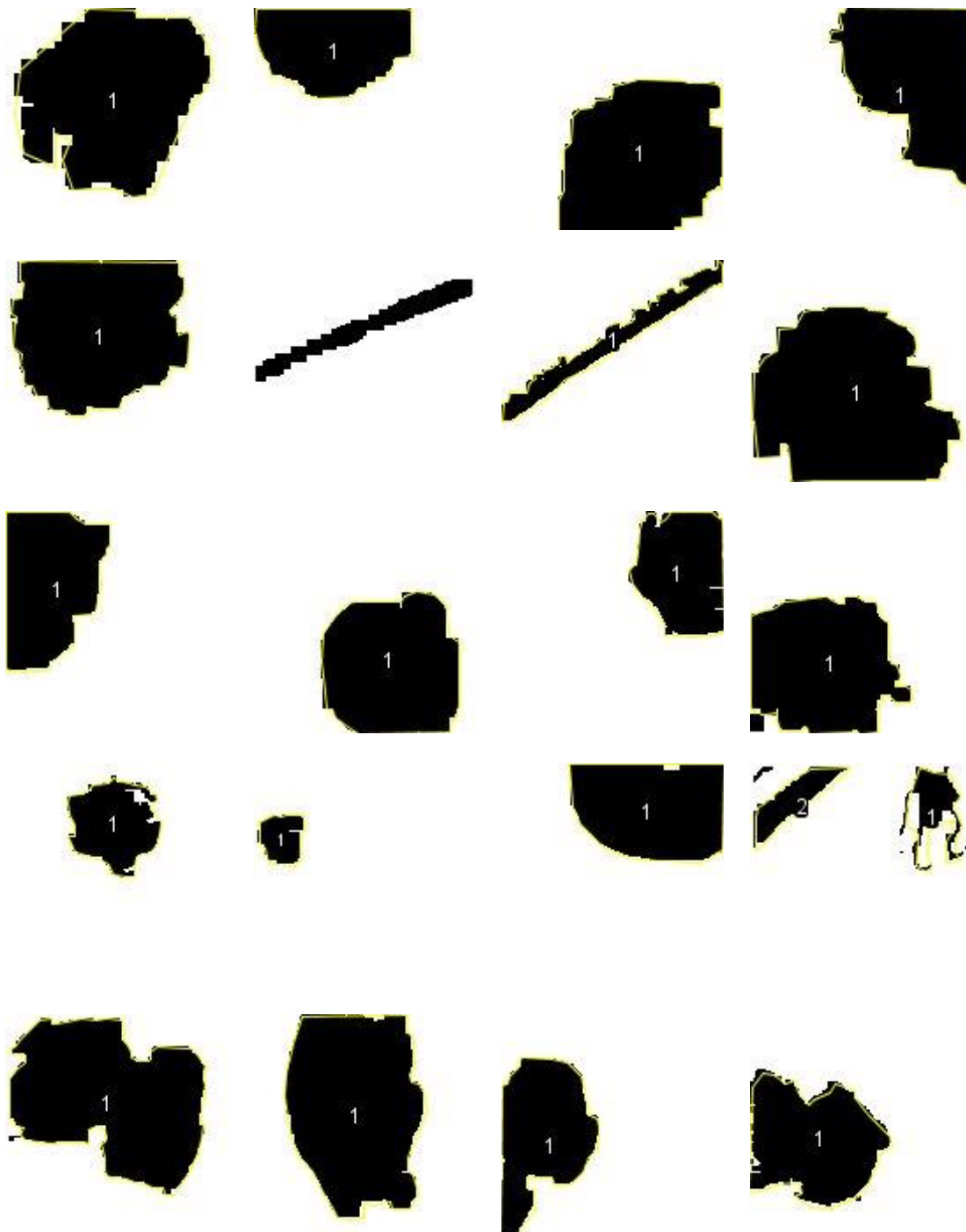


Figura 48: Conjunto de imágenes con extracción de características.

4.4 Resultados obtenidos al aplicar los clasificadores

Una vez que se realizó el procesamiento digital de imágenes y la extracción de características, las imágenes resultantes se utilizan para las dos últimas etapas de la metodología propuesta, que son el entrenamiento de los algoritmos y la clasificación de las imágenes celulares.

El entrenamiento de los algoritmos se realizó utilizando la metodología mencionada en la sección 3.5 y usando el 80 % de las fotografías de la base de datos.

Se clasificaron las imágenes usando los tres algoritmos supervisados de inteligencia artificial seleccionado para esta tesis, los cuales son Impulso adaptativo (Adaptative Boost), Bayes ingenuo (Naive Bayes) y Bosque aleatorio (Random Forest). Los resultados cuantitativos de la aplicación de cada algoritmo pueden ser visualizados en la matriz de confusión que se muestran en las Tablas 5, 6, 7, 8, 9 y 10, respectivamente. Mientras que una evaluación cualitativa del resultado se puede realizar, tomando como base el conjunto de 20 imágenes previamente utilizadas. Dichos resultados cualitativos se muestran para cada método en las Figuras 49, 50, y 51 respectivamente.

Tabla 5: Matriz de confusión del algoritmo impulso adaptativo.

Matriz de confusión								Adaptative Boost
a	b	c	d	e	f	Clasificado como	Color clase	
257123	1377	0	4934	0	1526	a = Célula		
1754	261339	0	891	60	352	b = Fondo		
0	0	264824	0	111	0	c = Lente negro		
4784	831	0	258637	7	994	d = Debris		
0	10	34	5	265546	1	e = Borde		
1431	317	0	1129	5	262094	f = Célula no viable		

Tabla 6: Métricas de desempeño del algoritmo de impulso adaptativo.

Parámetros						
TP	FP	Precisión	Sensibilidad	Media armónica	Área bajo la curva	Clase
0.970	0.006	0.970	0.970	0.970	0.983	a = Célula
0.988	0.002	0.990	0.988	0.989	0.994	b = Fondo
1.000	0.000	1.000	1.000	1.000	1.000	c = Lente negro
0.975	0.005	0.974	0.975	0.974	0.985	d = Debris
1.000	0.000	0.999	1.000	1.000	1.000	e = Borde
0.989	0.002	0.989	0.989	0.989	0.994	f = Célula no viable
0.987	0.003	0.987	0.987	0.987	0.993	Promedio

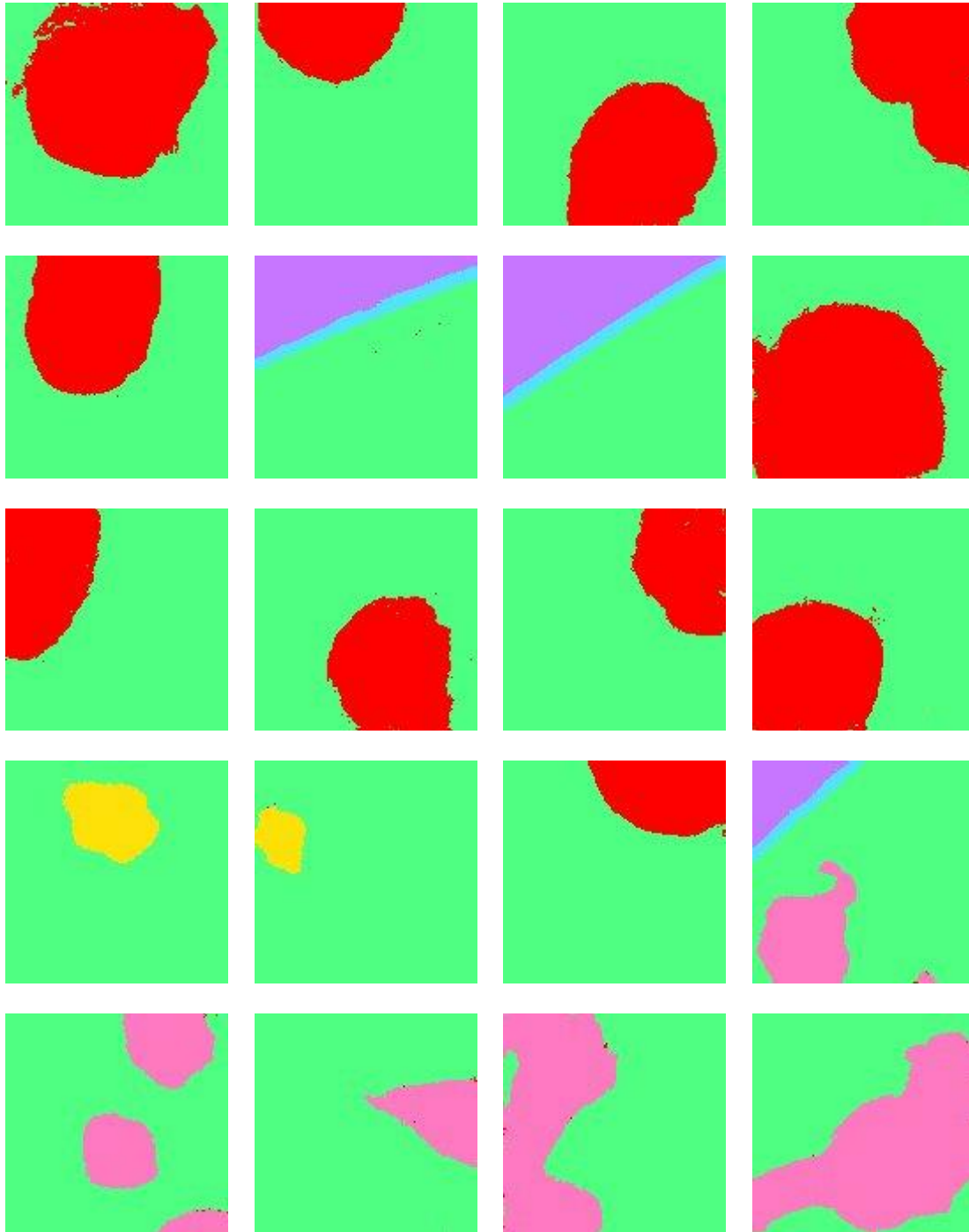


Figura 49: Conjunto de imágenes clasificadas, resultantes del algoritmo impulso adaptativo.

De la literatura especializada se derivaron métricas las cuales se reportan en la sección 3.7.1., dichas métricas son utilizadas para evaluar cuantitativamente el desempeño de los algoritmos. Como se puede observar en la matriz de confusión en la Tabla 5, los resultados arrojados por el algoritmo de impulso adaptativo fueron los siguientes: de un total de 1,590,116 de instancias clasificadas 1,569,563 fueron clasificadas correctamente, lo que corresponde al 98.7% de ella, lo que en consecuencia implica que se clasificaron incorrectamente 20,553 instancias, lo que corresponde al 1.29% de las instancias.

Se puede apreciar en la Figura, que en la matriz de confusión que la clase “Célula” es confundida por el algoritmo mayormente con la clase “Debris” y la clase “Célula no viable”.

Por otro lado, los parámetros de precisión, sensibilidad, media armónica, y área bajo la curva que se muestran en las columnas 3 a 6 de la Tabla 6 para cada una de las clases, presentan valores promedio cercanos a 1, lo que implica cuantitativamente que el proceso se está desarrollando correctamente. Puesto que, el parámetro precisión le brinda información al usuario sobre la calidad del modelo de clasificación entonces los resultados indican que el modelo de la clasificación es de buena calidad. Mientras que, el parámetro la sensibilidad indica la cantidad de instancias que el modelo es capaz de identificar, como su valor es cercano a 1, implica una porción marginal de instancias que no fueron identificadas correctamente.

Ahora, como la media armónica realiza una combinación de la precisión y la sensibilidad para obtener una sola medida promediable y como los dos parámetros anteriores obtuvieron buen resultado, entonces este indicador también presenta un buen desempeño. Por último, bajo la consideración de que el área bajo la curva permite apreciar el rendimiento del clasificador y por el resultado obtenido se puede indicar que rendimiento del clasificador es alto.

De la inspección cualitativa de los resultados obtenidos en la clasificación realizada por el algoritmo por parte de un experto en el reconocimiento de células viables de cáncer de mama y con experiencia en el trabajo de esta línea celular, se obtuvo una opinión satisfactoria del resultado. Una ilustración cualitativa de los resultados para el algoritmo de Bayes ingenuo se muestra en la Figura 50.

Tabla 7: Matriz de confusión del algoritmo de Bayes ingenuo.

Matriz de confusión								Naive Bayes
a	b	c	d	e	f	Clasificado como	Color clase	
263294	348	1	824	4	311	a = Célula		
478	263504	2	414	24	24	b = Fondo		
0	1	264800	2	92	123	c = Lente negro		
947	373	1	261228	9	2575	d = Debris		
4	4	18	6	265483	1	e = Borde		
539	31	115	2627	9	262210	f = Célula no viable		

Tabla 8: Métricas de desempeño del algoritmo de Bayes ingenuo.

Parámetros						
TP	FP	Precisión	Sensibilidad	Media armónica	Área bajo la curva	Clase
0.994	0.001	0.993	0.994	0.993	0.996	a = Célula
0.996	0.001	0.997	0.996	0.997	0.998	b = Fondo
0.999	0.000	0.999	0.999	0.999	1.000	c = Lente negro
0.985	0.003	0.985	0.985	0.985	0.991	d = Debris
1.000	0.000	0.999	1.000	1.000	1.000	e = Borde
0.987	0.002	0.989	0.987	0.988	0.993	f = Célula no viable
0.994	0.001	0.994	0.994	0.994	0.996	Promedio

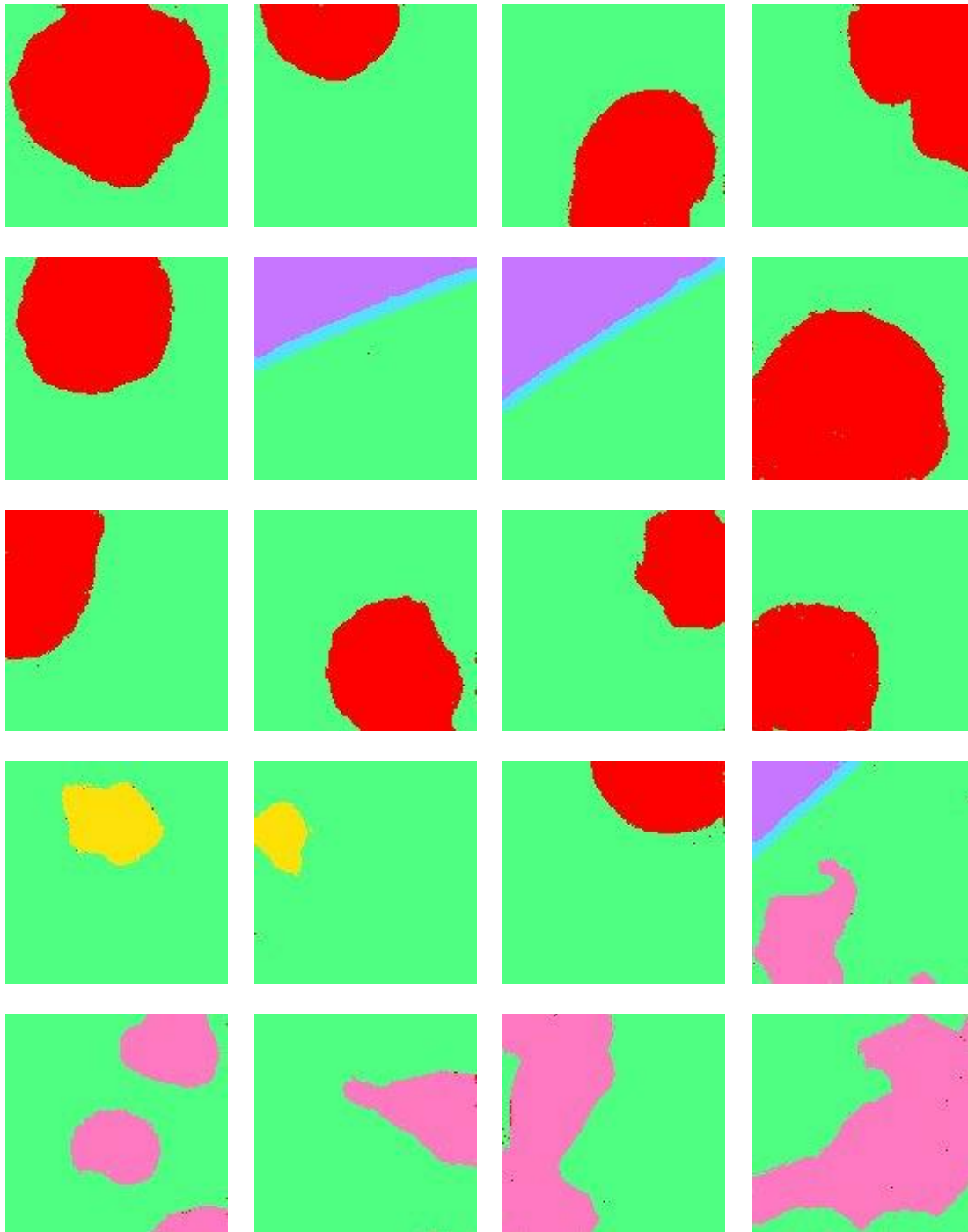


Figura 50: Conjunto de imágenes clasificadas, resultantes del algoritmo Bayes ingenuo.

Tabla 9: Matriz de confusión del algoritmo de bosque aleatorio.

Matriz de confusión							
a	b	c	d	e	f	Clasificado como	Color clase
261425	337	0	2411	41	818	a = Célula	
369	263857	0	336	35	218	b = Fondo	
0	0	264534	0	70	0	c = Lente negro	
2525	292	0	261974	12	597	d = Debris	
18	1	8	6	265057	0	e = Borde	
643	125	0	583	4	263753	f = Célula no viable	

Random Forest

Tabla 10: Métricas de desempeño del algoritmo de bosque aleatorio.

Parámetros						
TP	FP	Precisión	Sensibilidad	Media armónica	Área bajo la curva	Clase
0.989	0.003	0.987	0.986	0.986	0.993	a = Célula
0.996	0.001	0.997	0.996	0.997	0.998	b = Fondo
1.000	0.000	1.000	1.000	1.000	1.000	c = Lente negro
0.987	0.003	0.987	0.987	0.987	0.994	d = Debris
1.000	0.000	0.999	1.000	1.000	1.000	e = Borde
0.995	0.001	0.994	0.995	0.994	0.997	f = Célula no viable
0.994	0.001	0.994	0.994	0.994	0.997	Promedio

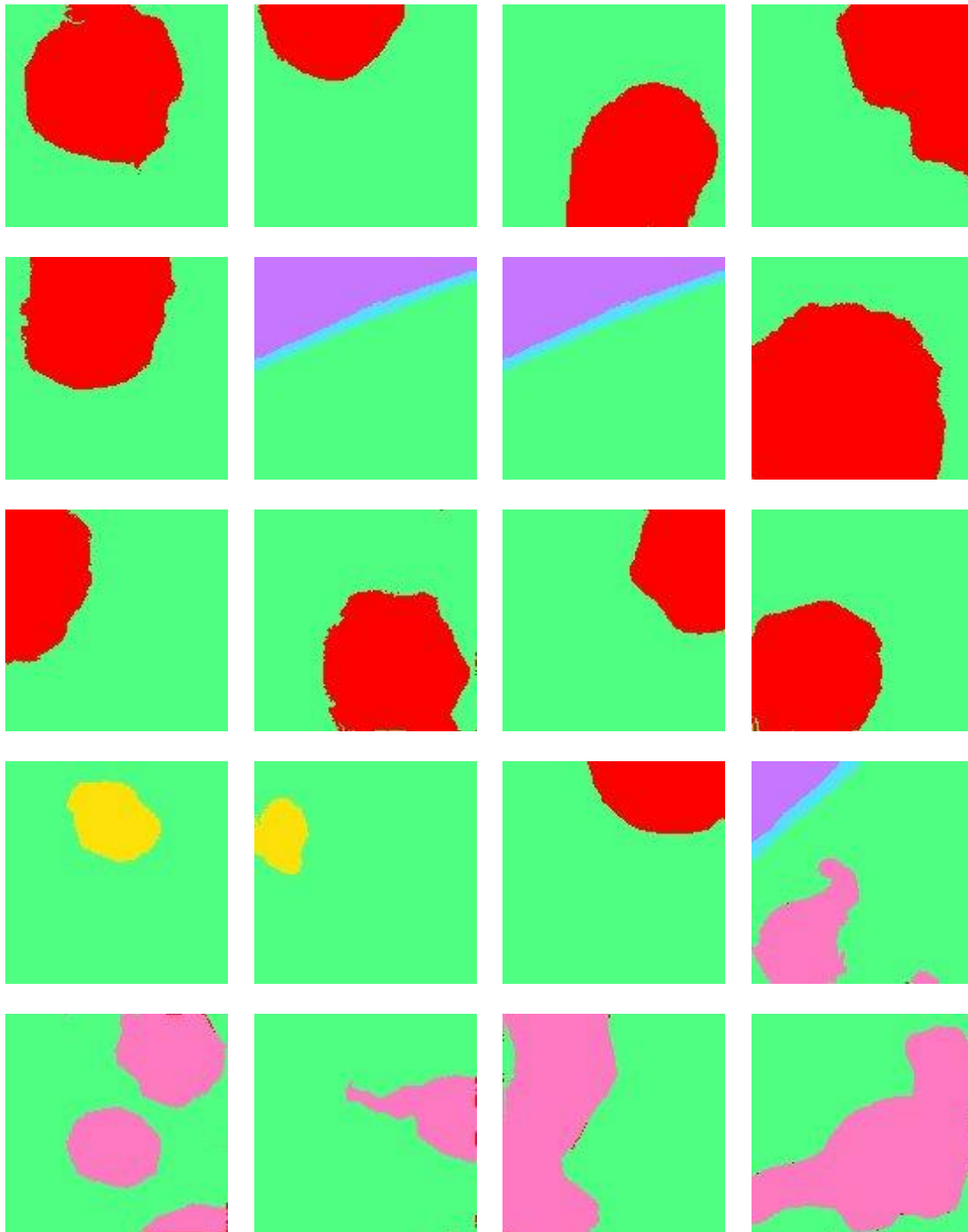


Figura 51: Conjunto de imágenes clasificadas, resultantes del algoritmo bosque aleatorio.

Los resultados arrojados por el algoritmo de Bosque aleatorio fueron los siguientes: de un total de 1,590,049 instancias clasificadas 1,580,600 fueron clasificadas correctamente, lo que corresponde al 99.4% y se clasificaron incorrectamente 9,449 instancias, lo que corresponde al 0.5%. Se aprecia en la matriz de confusión de la Tabla 9, que la clase “Célula” se confunde mayormente con la clase “Debris” y la clase “Célula muerta”.

4.5 Comparativo de los resultados de aplicar los clasificadores

Tabla 11: Tabla comparativa de los tres modelos de aprendizaje supervisado.

	TP	FP	Precisión	Sensibilidad	Media armónica	Área bajo la curva	% de I.C.	% de I.E.
Naive Bayes	0.994	0.001	0.994	0.994	0.994	0.996	99.3771%	0.6229%
Adaptative Boost	0.987	0.003	0.987	0.987	0.987	0.993	98.7075%	1.2925%
Random Forest	0.994	0.001	0.994	0.994	0.994	0.997	99.4057%	0.5943%

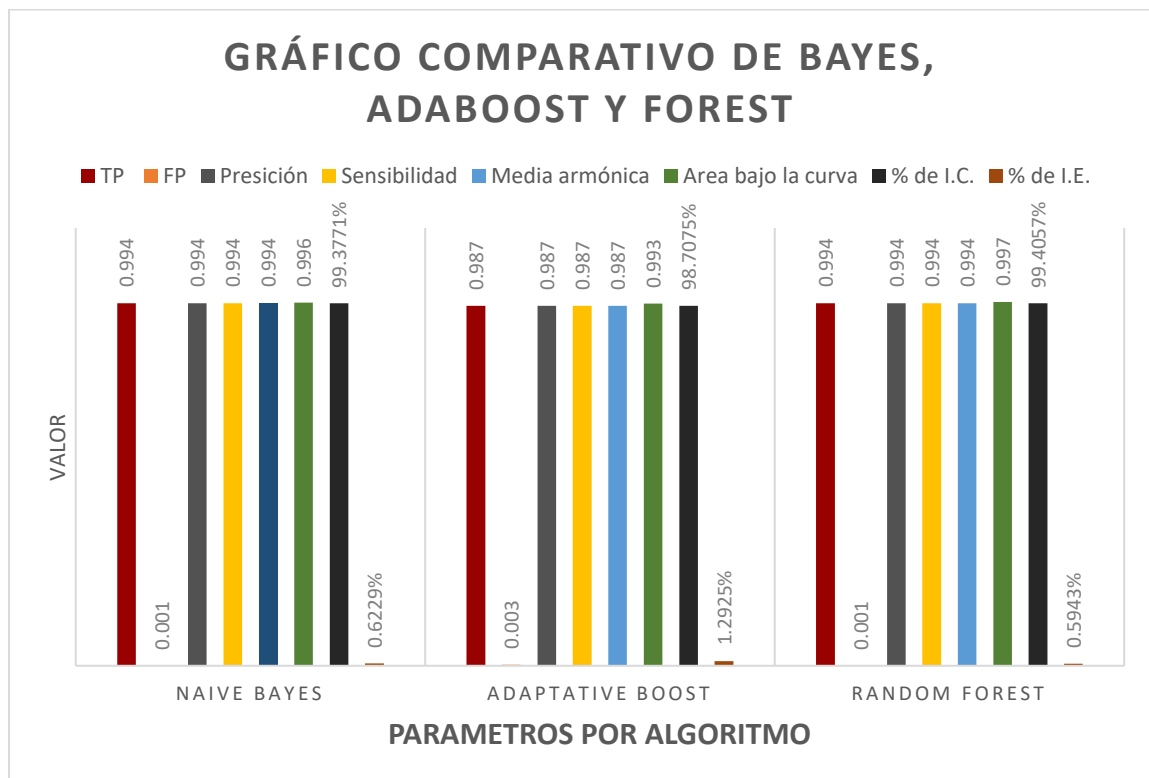


Figura 52: Gráfica comparativa de los tres algoritmos de la metodología propuesta.

En la Tabla 11 se presenta un resumen de las métricas de desempeño de los tres algoritmos de clasificación utilizados para identificar células viables de cáncer de mama en imágenes obtenidas al microscopio en condiciones de laboratorio, utilizando tres algoritmos de aprendizaje supervisado. Mientras que en la Figura 52 se muestra una gráfica de barras de los valores promedios obtenidos para cada métrica de desempeño. En ellas se puede observar que el algoritmo de bosque aleatorio, presenta un mejor desempeño que Bayes ingenuo e impulso adaptativo, se observa que existe una mayor calidad y rendimiento por parte del algoritmo de bosque aleatorio, con un 99.4% de instancias clasificadas correctamente, con lo que se puede observar que la metodología propuesta fue llevada a cabo exitosamente.

Capítulo 5

5. Conclusiones

En la evaluación de fármacos, los ensayos de citotoxicidad *in vitro* forman parte de la primera etapa. Con base en los resultados de la evaluación citotóxica es posible obtener información de la biocompatibilidad del fármaco. Para ello, es necesario conocer la viabilidad celular que forma parte del cultivo o de la línea celular con el objetivo de mantener una proliferación celular adecuada en las muestras usadas en la evaluación. El proceso de determinar la viabilidad celular en ambiente de laboratorio implica un nivel inaceptable de subjetividad, que es una consecuencia del proceso tedioso ante el cual el experto se enfrenta.

En este trabajo de tesis, se presentan los resultados de un estudio comparativo de tres algoritmos de inteligencia artificial supervisados entrenados en la identificación de células viables de cáncer de mama en imágenes obtenidas al microscopio en ambiente de laboratorio, de línea celular MDA-MB-231. Dichos resultados pueden ser utilizados para el desarrollo de un sistema de control de calidad que le permita monitorear la viabilidad de los cultivos, a bajo costo y con un menor consumo de tiempo.

Se presenta una metodología para evaluación de proceso de clasificación de la viabilidad de células de cáncer de mama con base imágenes tomadas al microscopio, obtenidas en ambiente de laboratorio. La cual utiliza con éxito técnicas de preprocesamiento digital de imágenes, tales como el filtro Kuwahara, el filtro de difusión anisotrópica y el filtro de agudizado, para obtener imágenes con una menor proporción de ruido, un fondo uniforme y contornos mejor definidos de las células. Esto permitió hacer eficiente el proceso de segmentación en un grupo de 42,000 imágenes.

Se han empleado técnicas de segmentación para procesamiento digital de imágenes, tales mediante el uso operador gradiente de Sobel y la umbralización a fin de distinguir las células del fondo.

Se han utilizado parámetros morfológicos con el objetivo de analizar las formas de las células y extraer las características de estas, tales como el tamaño, su forma, y textura. Lo que permitió contar con información cuantificable auxiliar en la discriminación de las células viables y no viables en el proceso de etiquetado por parte de un experto en el reconocimiento de células viables de cáncer de mama y con experiencia en el trabajo de esta línea celular. Las etiquetas fueron asignadas durante el proceso de entrenamiento, esto fue realizado mediante la interfaz gráfica de Segmentación Entrenable de Weka.

Con base a la metodología y usando los softwares libres FIJI® y WEKA®, se ha evaluado el desempeño de los algoritmos de aprendizaje supervisado: Impulso adaptativo, Bayes ingenuo, y Bosque aleatorio para diferenciar las células de cáncer de mama viables de otras cinco clases.

Finalmente, de acuerdo con los resultados obtenidos en las matrices de confusión y utilizando métricas de desempeño se concluye que el algoritmo de inteligencia artificial que discrimina con una mayor tasa de instancias clasificadas correctamente en la identificación de células viables de cáncer de mama de línea celular MDA-MB-231 es el de Bosque aleatorio.

Cada uno de los tres algoritmos evaluados conforme a la metodología propuesta, fueron entrenados con 1,590,000 instancias de entrada del software WEKA® para procesamiento de imágenes.

5.1 Líneas futuras de investigación

Es posible realizar una modificación en la etapa de extracción de características y en el entrenamiento del modelo de aprendizaje, para detectar heridas en el cultivo celular, que es otra clase de estudio realizado en laboratorio.

La etapa de clasificación es un área en la cual podría seguirse explorando, para ello se puede comparar el algoritmo que obtuvo una mejor tasa de instancias clasificadas correctamente, que en este caso es bosque aleatorio con redes neuronales u otros algoritmos.

La metodología propuesta en el presente trabajo de tesis podría aplicar en encontrar microcalcificaciones en placas tomadas obtenidas en mamografías, esto en aplicaciones en imágenes médicas.

APENDICE

Preprocesamiento.

- Aplicar los filtros espaciales: Difusión anisotrópica, Agudizado, Kuwahara.

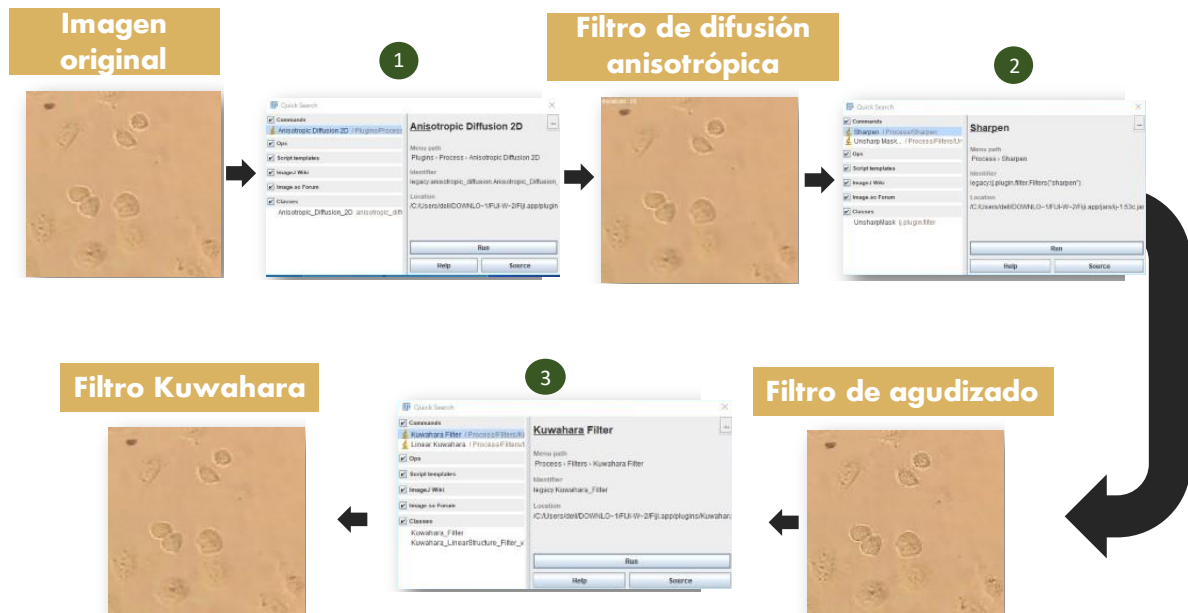


Figura 53: Método para aplicar filtros espaciales en FIJI®.

Detección de objetos.

- Aplicar el operador gradiente de Sobel, con el propósito de obtener los bordes de las células.

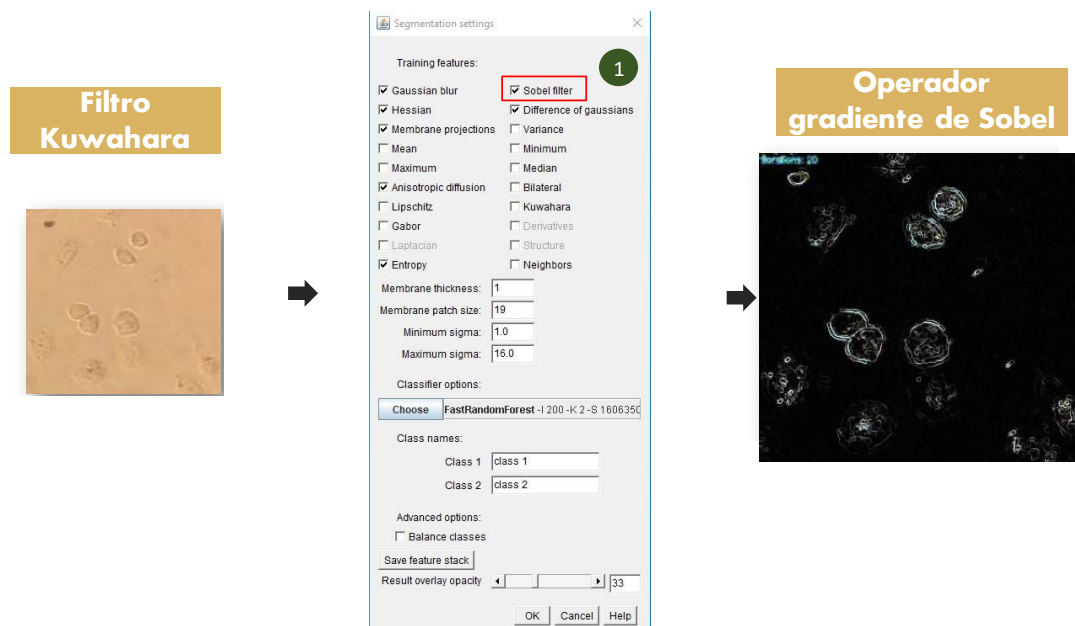


Figura 54: Método para aplicar el operador gradiente de Sobel en TWS.

Extracción de características

- ConFigurar las características que serán extraídas, mediante la ventanilla de “establecer medidas”.
- Extraer las características de la imagen, mediante la ventanilla de “analizar partículas”.



Figura 55: Método para extraer las características morfológicas en FIJI®.

Entrenamiento.

- Inicializar Segmentación Entrenable de Weka.
- Asignar etiquetas a los objetos de interés en la imagen (Células viables, células no viables, debris, fondo de caja de Petri).

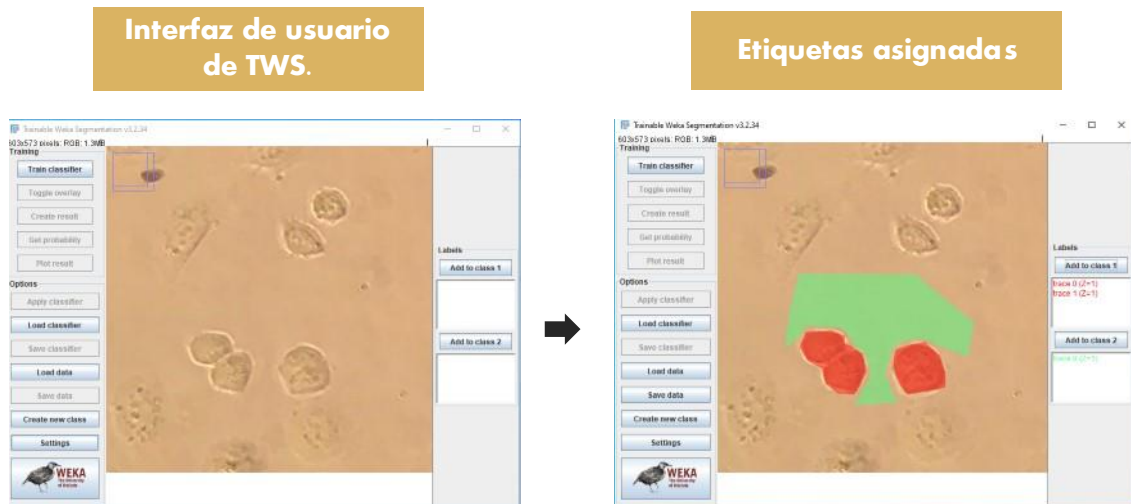


Figura 56: Método para inicializar TWS y generar etiquetas.

- Asignar una clase al conjunto de etiquetas que se realizaron previamente.
- Cargar el algoritmo deseado.
- ConFigurar el algoritmo con los parámetros deseados.

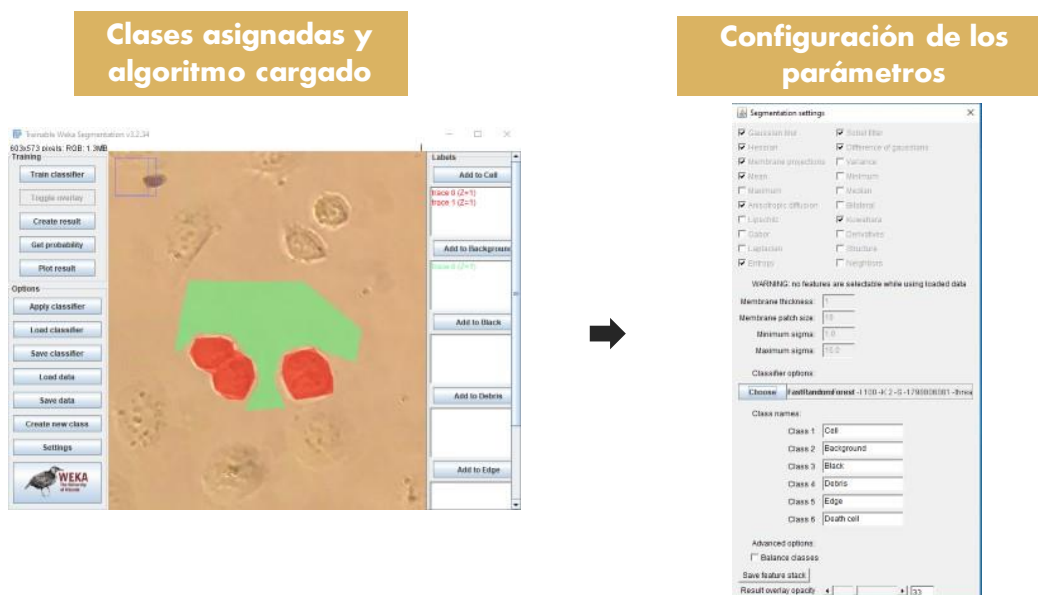


Figura 57: Método para asignar clases y configurar el algoritmo de I.A.

- Entrenar el algoritmo en base a las clases asignadas.

Entrenando algoritmo

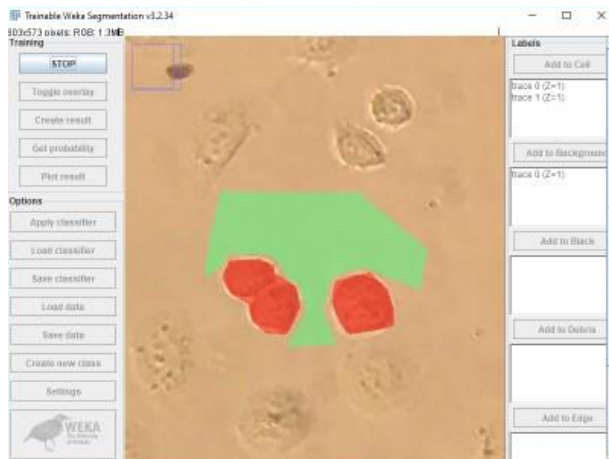


Imagen obtenida después
del entrenamiento

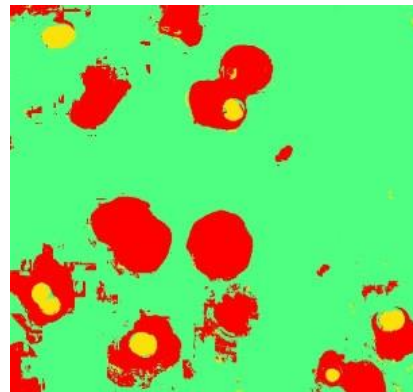


Figura 58: Método para entrenar algoritmo en TWS, e imagen resultante del entrenamiento.

Clasificación.

- Cargar el modelo supervisado de inteligencia artificial, el archivo “. model”.
- Cargar los vectores de datos con los atributos e instancias, el archivo “. arff”.
- Aplicar el clasificador en una o más imágenes.

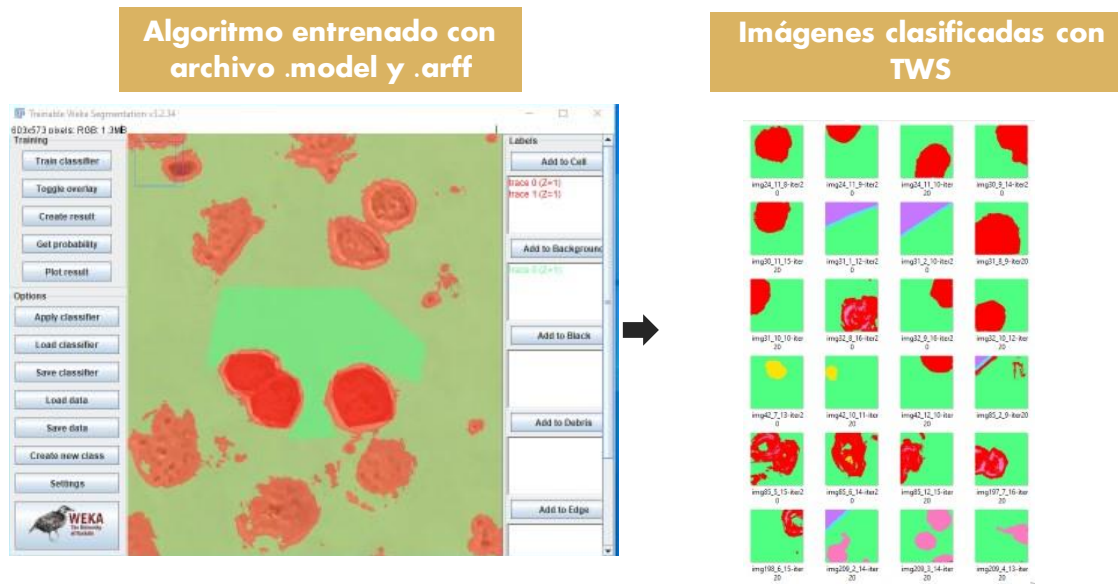


Figura 59: Método para clasificar múltiples imágenes en TWS.

Evaluación

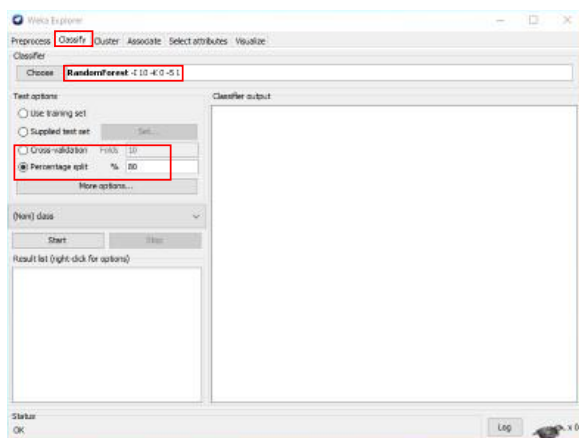
- Inicializar WEKA®.
- Inicializar la aplicación de “explorador” (explorer) del menú de WEKA®.
- Seleccionar la pestaña de “preprocesar” (preprocess).
- Abrir el conjunto de datos, con extensión “. arff”.



Figura 60: Método para visualizar los vectores de datos en WEKA®.

- Seleccionar la pestaña de “clasificar”.
- Seleccionar el modelo de aprendizaje supervisado.
- Indicar la partición de datos.
- Obtener la matriz de confusión.
- Comparar las métricas y la matriz de confusión con los resultados de los algoritmos restantes.

Pestaña clasificar



Matriz obtenida

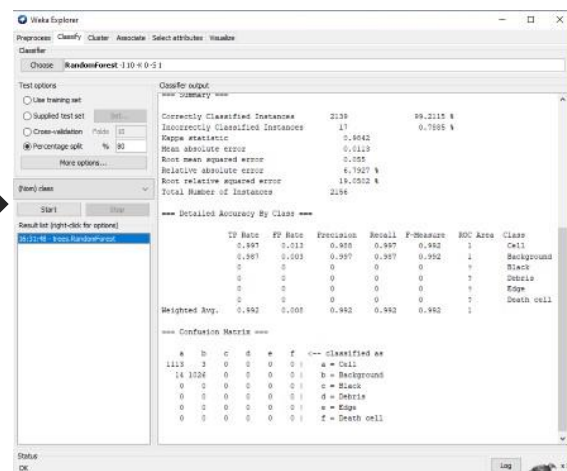


Figura 61: Método para obtener la matriz de confusión y sus métricas en WEKA®.

Referencias bibliográficas

- [1] Ríos Hernández, Miriam, & Cepero Cañas, Janet (2006). Citotoxicidad in vitro: sistema para la evaluación de biomateriales y equipos médicos implantables en Cuba. Revista CENIC. Ciencias Biológicas, 37(3),173-176. ISSN: 0253-5688. Disponible en: <https://www.redalyc.org/articulo.oa?id=181220529011>
- [2] Stoddart M. Mammalian Cell viability: Methods and Protocols, Methods in Molecular Biology. En: Stoddart M. Cell Viability Assays: Introduction. Davos Platz: Suiza; 2011. p. 1-6.
- [3] LOPEZ ALVAREZ, E. S. T. R. E. L. L. A. M. A. R. C. E. L. A. (2019). MÉTODOS PARA DETERMINAR LA VIABILIDAD CELULAR CON APLICACIÓN EN ODONTOLOGÍA. Universidad de San Martín de Porres.
- [4] Quintanilla Domínguez, J. (2015, octubre). Aplicación de técnicas de inteligencia artificial para contribuir en la detección de microcalcificaciones en mamografía digitalizada (TFM). Aplicación de técnicas de inteligencia artificial para contribuir en la detección de microcalcificaciones en mamografía digitalizada. <https://oa.upm.es/35038>.
- [5] Sommer, C., & Gerlich, D. W. (2013). Machine learning in cell biology – teaching computers to recognize phenotypes. Journal of Cell Science. <https://doi.org/10.1242/jcs.123604>
- [6] Gonzalez, R. C. (2022b). Digital Image Processing, Global Edition (4th edition). Pearson.
- [7] Tschumperle, D., & Deriche, R. (2005). Vector-valued image regularization with PDEs: a common framework for different applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(4), 506–517. <https://doi.org/10.1109/tpami.2005.87>

- [8] Pulido López, D. A. (2018, junio). Sobre el modelo de difusión anisotrópica de Perona-Malik (TFM). Universidad Nacional de Colombia. <https://repositorio.unal.edu.co/handle/unal/63516>
- [9] Hassan, N. M. H., Universidad de Sevilla. Departamento de Electrónica y Electromagnetismo, & Universidad de Sevilla. Facultad de Física. (2009). Diseño de circuitos para tratamiento de imágenes aplicando técnicas basadas en soft computing. Universidad de Sevilla. ISBN: 9788469277607
- [10] Gonzalez, R.C., Wintz, P. (1996), Procesamiento digital de imágenes. Addison Wesley.
- [11] Introduction. (2022). FIJI® ImageJ ®. <https://imagej.nih.gov/ij/docs/intro.html>
- [12] Ferreira T and Rasband WS. “ImageJ User Guide — IJ 1.46”, imagej.nih.gov/ij/docs/guide/, 2010–2012
- [13] Ignacio Arganda-Carreras, Verena Kaynig, Curtis Rueden, Kevin W Eliceiri, Johannes Schindelin, Albert Cardona, H Sebastian Seung, Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification, Bioinformatics, Volume 33, Issue 15, 01 August 2017, Pages 2424–2426, <https://doi.org/10.1093/bioinformatics/btx180>
- [14] Trainable Weka Segmentation. (2022). ImageJ Wiki. <https://imagej.net/plugins/tws/>
- [15] McCarthy et al. (31, agosto de 1955) “A Proposal for the Dartmouth Summer Research project on Artificial Intelligence”. Universidad de Stanford Recuperado de: <https://web.archive.org/web/20080930164306/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>.
- [16] Gamarra Moreno, A. H. (2020). Modelos de aprendizaje automático basado en técnicas supervisadas para la predicción de la mortalidad perinatal en la Región Junín (TFM). Universidad Nacional Federico Villarreal. <http://repositorio.unfv.edu.pe/handle/UNFV/4355>
- [17] [Rish, I. (2001). An empirical study of the naive Bayes classifier. IBM Research, Report, Yorktown Heights.

- [18] Alarcón Jaimes, C. (2015). OPTIMIZACIÓN DEL CLASIFICADOR “NAIVE BAYES” USANDO ÁRBOL DE DECISIÓN C4.5 (TFM). Universidad Nacional Mayor de San Marcos. <https://cybertesis.unmsm.edu.pe/handle/20.500.12672/4183>
- [19] Alarcón Flores, J. B. (2017). Modelos de minería de datos: random forest y adaboost, para identificar los factores asociados al uso de las TIC (internet, telefonía Fija y televisión de paga) en los hogares del Perú. 2014 (TFM). Universidad Nacional Mayor de San Marcos. <https://cybertesis.unmsm.edu.pe/handle/20.500.12672/7404>
- [20] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall Geoffrey Holmes, and Sally Jo Cunningham. WEKA®: Practical machine learning tools and techniques with java implementations. Department of Computer Science. University of Waikato. New Zealand. <http://www.cs.waikato.ac.nz/~ml/publications/1999/99IHW-EF-LT-MH-GH-SJC%-Tools-Java.pdf>.
- [21] Anuranjeeta, A., K. Shukla, K., Tiwari, A., & Sharma, S. (2017). Classification of Histopathological Images of Breast Cancerous and Non Cancerous Cells based on Morphological Features. Biomedical and Pharmacology Journal, 10(1), 353–366. <https://doi.org/10.13005/bpj/1116>
- [22] Arora, S., & Barak, B. (2009). Computational complexity: A modern approach.
- [23] Espinosa-Zúñiga, Javier Jesús. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería, investigación y tecnología, 21(3), 00002. Epub 02 de diciembre de 2020. <https://doi.org/10.22201/fi.25940732e.2020.21.3.022>
- [24] Mateo Jiménez, M. (2020, agosto). Reconocimiento facial como medida de seguridad para alertar el robo de automóviles (TFM). Universidad Autónoma del Estado de México. https://ri.uaemex.mx/bitstream/handle/20.500.11799/109601/Tesis_Mayra%20Mateo_vF%20sin%20oficios.pdf?sequence=1&isAllowed=y

[25] Lenz, D., Gasparini, L., Macedo, N., Pimentel, E., Fronza, M., Junior, V., Borges, W., Cole, E., Andrade, T., & Endringer, D. (2017). In vitro cell viability by CellProfiler® software as equivalent to MTT assay. *Pharmacognosy Magazine*, 13(50), 365. <https://doi.org/10.4103/0973-1296.210176>