Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Clasificación Automática de Células Mediante Análisis de Texturas

Tesis Profesional

Que para obtener el título de:
Licenciada en Ciencias de la Computación

Presenta:
Dolores Limón Romero

Asesores:
Dra. Irene Olaya Ayaquica Martínez
Dr. Rafael Lemuz López

Octubre 2013

Agradecimientos

Agradezco al apoyo brindado por PROMEP al haberme otorgado la beca de tesis durante el año de realización del proyecto de investigación: *Desarrollo de Software para el Hospital de Habilidades y Destrezas*, con folio BUAP-PTC-255 y número de convenio 103.5/11/4476.

Índice general

Índice de figuras

Índice de cuadros

Resumen

En este trabajo se propone un algoritmo para la clasificación automática de células. El algoritmo se basa en el uso de descriptores de textura que se obtienen a partir de los niveles de intensidad de las imágenes, tales descriptores de textura son obtenidos mediante el método estadístico de la Matriz de Co-ocurrencia de Niveles de Gris (GLCM - Gray Level Coocurrence Matrix) y la Transformada Discreta Wavelet Haar 2D (DWT - Discrete Wavelet Transform) que es una función matemática utilizada para analizar una señal dependiente del tiempo a diferentes resoluciones. Describiremos brevemente los fundamentos y profundizaremos en el proceso de la extracción de características en ambos métodos.

Se plantea una breve introducción a la clasificación, metodologías y tipos de clasificación; en este caso la Clasificación Supervisada ya que se conocen los tipos de células al que pertenece el grupo de imágenes. Se describen brevemente tres algoritmos de clasificación automática: Árboles de Regresión y Clasificación (CART - Classification And Regression Trees), Redes Neuronales Probabilísticas (PNN - Probabilistic Neural Network) y Máquinas de Soporte Vectorial (SVMs - Support Vector Machines).

Para este trabajo la fase de clasificación se realizó utilizando Árboles de Regresión y Clasificación, mostramos la evaluación del algoritmo utilizando diferentes conjuntos de imágenes de prueba desde 300 a 700. El algoritmo integrará modelos matemáticos y computacionales de las áreas del reconocimiento de patrones y el análisis de imágenes.

Capítulo 1

Introducción

Contenido

1.1. Motivación

La división de células juega un rol crítico en el estudio de enfermedades y el desarrollo de nuevos fármacos. El análisis del fenotipo de división de células en alto contenido de proyección basado en imágenes microscópicas se basa en la segmentación nuclear automatizada y clasificación de fases del ciclo celular. Una identificación automatizada de la fase del ciclo celular ayuda a los biólogos a cuantificar el efecto de perturbaciones genéricas y los tratamientos con nuevos medicamentos. En este trabajo seguiremos una línea de investigación reciente que ha tratado con el problema de reconocer automáticamente el ciclo celular utilizando imágenes 2D de células.

En años recientes, las metodologías de visión por computadora han sido aplicadas en los campos de la informática de la salud y telemedicina para ayudar en el diagnóstico automático de enfermedades. La enorme colección de imágenes médicas generada todos los días alrededor del mundo han ayudado al interés en el diagnóstico de salud automatizado. Por ejemplo, los departamentos de radiología especializados produce más de 12,000 imágenes al día [1]. Por otra parte, medidas de diagnósticos automatizados han mostrado gran potencial para reducir errores de diagnóstico. Uno de los más importantes predictores del rendimiento del cuidado de la salud es la precisión y eficiencia en los diagnósticos médicos. Un estudio en Harvard reportó que el error de diagnóstico tiene un impacto negativo sustancial en el cuidado del paciente, tal como un costo incremental por paciente de $4,685 USD y un incremento en la duración media de la estancia por 4.6 días [2]. Sólo en Estados Unidos, errores médicos resultan en 44,000 a 98,000 muertes innecesarias cada año y 1 millón de lesiones en exceso [3]. Irónicamente, la mayoría de los errores de diagnóstico son evitables.

Investigaciones muestran que los errores de diagnóstico a menudo ocurren cuando los médicos son inexpertos y nuevos procedimientos son introducidos. Además, se ha encontrado que la edad, la atención compleja, la atención de urgencia y la estancia prolongada en el hospital se correlacionan con errores de diagnóstico. Los dos tipos de errores de diagnóstico incluyen errores por omisión, o fallas de acción tal como un diagnóstico perdido o una evaluación tardía, y errores de comisión, o acciones incorrectas tal como administrar el medicamento incorrecto al paciente equivocado en el tiempo equivocado.

La aplicación de sistemas de información automatizados en el análisis médico ha brindado una gran esperanza en la reducción de ambos tipos de errores basados en humanos[4].

1.2. Objetivos

1.2.1. General

Desarrollar un algoritmo que permita clasificar automáticamente células de acuerdo a los patrones que se definen en un proceso de tinción.

1.2.2. Particulares

- Desarrollar un algoritmo de procesamiento digital de imágenes para estandarizar el nivel de contraste de las imágenes.
- Desarrollar un algoritmo para definir un vector de características distintivas de las imágenes.
- Implementar un algoritmo de clasificación automática basado en técnicas de aprendizaje supervisado.

1.3. Hipótesis de Investigación

Es posible caracterizar las imágenes celulares para su clasificación utilizando descriptores de textura.

1.4. Revisión del estado del arte

Las divisiones de células y su regulación juegan un rol importante en enfermedades y desarrollo de nuevos medicamentos. El ciclo celular puede ser dividido en dos periodos principales: interfase y mitosis. Durante la interfase las células crecen, duplican su ADN y acumulan nutrientes y productos de genes necesarios para la mitosis. Durante la mitosis, las mismas células se dividen y dividen el ADN genómico entre las dos células hijas. La mitosis puede ser adicionalmente subdividida dentro de varias fases distintas: profase, metafase, anafase y telofase. Las fases de las células pueden ser identificadas por su apariencia en imágenes de microscopio de alta resolución. La figura 1.1 muestra ejemplos de la típica apariencia del marcador de cromatina histona-GFP (Green Flourescent Protein) en diferentes fases del ciclo celular.

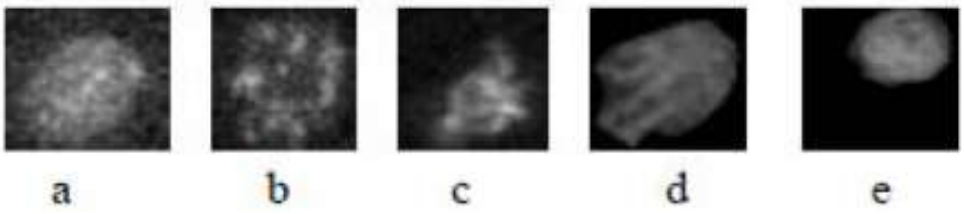


Figura 1.1: Proyecciones de máxima intensidad de los núcleos en un embrión vivo Drosophila etiquetados con la histona-GFP (a) Interfase, (b) Profase, (c) Metafase, (d) Anafase y (e) Telofase.

La clasificación automática de la fase celular es un paso esencial en el análisis de imágenes que utiliza alta resolución al procesar grandes poblaciones de células que permite la cuantificación de la progresión del ciclo celular, lo que es muy importante para el desarrollo de la biología, el estudio celular de cáncer y el descubrimiento de fármacos. Por ejemplo, midiendo la duración de fases individuales del ciclo celular bajo diferentes condiciones genéticas y tratamientos médicos se puede incrementar la comprensión de los mecanismos biológicos en enfermedades oncológicas y mejorar la eficacia del descubrimiento de nuevos fármacos y su desarrollo [5].

La clasificación de la fase celular es crucial para las pantallas de alta resolución basadas en imágenes, tales como el proyecto Mitocheck que están dirigidas a la identificación y caracterización de genes implicados en la división celular [6]. Varios grupos de investigación Bioimagen han abordado este problema difícil. La mayoría de los estudios incluyeron imágenes 2D. En un estudio se trató con imágenes 3D, pero las características celulares se extrajeron a partir del segmento con más información [7]. Las características dinámicas han sido ampliamente utilizadas para la clasificación de la fase celular [8], sin embargo como se ha reportado en varias investigaciones, los algoritmos de seguimiento tienden a ser menos fiables y el contexto de la información se hace menos confiable cuando las células están densamente pobladas y/o se mueven a gran velocidad.

En los últimos años, la microscopía confocal de escaneo láser (CLSM) se ha convertido en una modalidad común de imagen para visualizar en 3D las células marcadas con tintes fluorescentes. La dimensión extra en comparación con la microscopia convencional 2D promete mejorar la comprensión de los mecanismos bio-moleculares. Otra aplicación de la identificación automatizada de la fase del ciclo celular es la de mejorar el rastreo celular mediante el análisis de imágenes en el tiempo. En los tejidos vivos, las células pueden desplazarse grandes distancias. Desplazamientos significativos en períodos cortos (por ejemplo, un minuto) son especialmente pronunciados en la mitosis de embriones de Drosophila. Desde que ocurren las fases del ciclo celular en un orden fijo, el seguimiento puede ser mejorado usando este conocimiento biológico previamente. Por lo tanto, es esencial para desarrollar un algoritmo de clasificación de fase celular que utilice la información de la imagen 3D y no se base en rasgos dinámicos obtenidos por rastreo celular.

1.5. Organización del documento

El documento de tesis se encuentra organizado de la siguiente manera: En el capítulo 1 presentamos una introducción al trabajo de la tesis, describiendo los objetivos y el trabajo relacionado. En el capítulo 2 se hace una revisión del tema de representación de las texturas en imágenes y algunas métricas de caracterización. En el capítulo 3 describimos los conceptos del aprendizaje supervisado que dan sustento al trabajo de investigación. En el capítulo 4 se puede consultar el desarrollo del trabajo realizado y los resultados obtenidos. Finalmente en el capítulo 5 se discuten las conclusiones y el trabajo futuro.

Capítulo 2

Análisis de Texturas

Contenido

En este trabajo utilizaremos un enfoque muy similar al propuesto en [9] donde se propone un sistema automatizado de reconocimiento de patologías cutáneas humanas, analiza la textura de imágenes de la piel mediante técnicas de reconocimiento basadas en textura con imágenes en niveles de gris. Este trabajo, de forma similar, utiliza la matriz de co-ocurrencia y la descomposición wavelet de las imágenes como características para discriminar entre las distintas etapas del ciclo celular. El análisis de textura es uno de los aspectos fundamentales de la visión humana que permite discriminar entre las superficies y objetos.

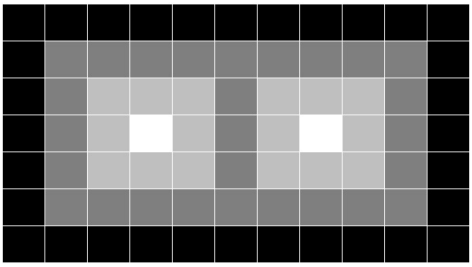
En el campo del procesamiento digital de imágenes, las técnicas de visión por computadora pueden tomar ventaja de las señales proporcionadas por textura de superficie para distinguir y reconocer objetos.

La textura se refiere a los patrones visuales o arreglo espacial de los píxeles que se obtienen a partir de la intensidad o el color y tiene la propiedad de describir de forma más completa ciertos patrones de las células [10]. Los investigadores han propuesto numerosas metodologías para analizar y reconocer automáticamente texturas. Uno de los primeros estudios involucrados es la derivación de las medidas de textura de la energía que utilizan un conjunto de máscaras simples (vertical, horizontal, diagonal y antidiagonal), para el uso de filtros de Gabor en varias aplicaciones de análisis de imágenes, incluyendo la clasificación de textura y segmentación.

En este trabajo utilizaremos el análisis de texturas a partir de una colección de imágenes médicas para la clasificación automática del ciclo celular. Este enfoque usa características basadas en una matriz de co-ocurrencia de nivel de gris (GLCM) y una matriz de descomposición Wavelet (WDM - Wavelet Decomposition Matrix).

2.1. Matriz de Co-ocurrencia

Introducido por Robert Haralick [11], una GLCM es un método estadístico popular para en análisis de textura. Una GLCM indica la probabilidad de que el nivel de gris i ocurra en la zona de nivel de gris j a una distancia d en la dirección θ. Algunas GLCM pueden ser calculadas desde imágenes de texturas usando diferentes valores de d y θ, y estos valores de probabilidad crean la matriz de co-ocurrencia $G(i, j|d, \theta)$. Por ejemplo: considerando una sección de una imagen I de 11 $\times$ 7 pixeles teniendo además 4 intensidades de niveles de gris (ver figura 2.1).



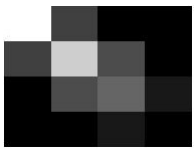
$$\text{I}= \begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 2 & 2 & 1 & 2 & 2 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 2 & 1 & 2 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 2 & 2 & 1 & 2 & 2 & 2 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Figura 2.1: GLCM Representación de la imagen en niveles de grises y su matriz correspondiente.

Para calcular la frecuencia de un tono gris alrededor de otros, formamos una matriz 4 $\times$ 4 (debido a que son cuatro tonos de gris distintos), indicando las frecuencias con números secuenciales a lo largo de la izquierda (pixel de referencia) y superior (pixel vecino). Podemos entonces calcular las frecuencias con la que cada par (referencia, vecino) de tonos de gris ocurren juntos en la imagen I.

Es decir, para un tono de referencia gris i, ¿Cuántas veces el vecino de tono gris j se produce cerca de él dentro de I? esto constituye el (i, j)-ésimo elemento de la matriz G de GLCM:

$$G = \begin{bmatrix} 20 & 5 & 0 & 0 \\ 5 & 16 & 6 & 0 \\ 0 & 6 & 8 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$



Para simplificar los cálculos, consideramos la distancia d como 1 (sólo considerando pixeles adyacentes) y el ángulo θ como 0°(a lo largo del eje positivo de X de izquierda a derecha). Por ejemplo, 0 (pixel de referencia con intensidad 0) adyacente a 0 (pixel vecino con intensidad 0) en I ocurre 20 veces (filas 1 y 7), así que colocamos 20 en la posición (0,0) de G. De igual manera, 0 adyacente a 1 ocurre cinco veces (filas 2, 3, 4, 5 y 6), por lo que la posición (0,1) contiene un 5; y 0 adyacente a 2 no ocurre, y la posición (0,2) contiene 0 ocurrencias, y así sucesivamente. Este procedimiento se repite para todos los pares de intensidades.

Si nos movemos a lo largo del eje negativo de X, esto es, si hubiéramos visto de derecha a izquierda, entonces la matriz formada debería ser la transpuesta de la matriz G^T. Para hacer la matriz independiente de este factor, la transpuesta se suma a la matriz original para hacerla simétrica $S = G + G^T$:

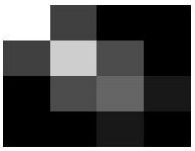
$$G + G^T = \begin{bmatrix} 20 & 5 & 0 & 0 \\ 5 & 16 & 6 & 0 \\ 0 & 6 & 8 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix} + \begin{bmatrix} 20 & 5 & 0 & 0 \\ 5 & 16 & 6 & 0 \\ 0 & 6 & 8 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

$$G + G^T = \begin{bmatrix} 40 & 10 & 0 & 0 \\ 10 & 32 & 12 & 0 \\ 0 & 12 & 16 & 4 \\ 0 & 0 & 4 & 0 \end{bmatrix} = S$$

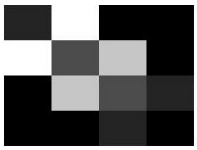
Finalmente normalizamos la matriz GLCM simétrica dividiendo cada elemento por la suma de todos los elementos para formar S_0.

El subíndice 0 indica el ángulo $\theta = 0°$, por ejemplo dirección horizontal. También podemos calcular GLCMs a lo largo de otras tres direcciones: vertical ($\theta = 90°$), diagonal derecha ($\theta = 45°$), y diagonal izquierda($\theta = 135°$), generando matrices S_{45}, S_{90}, y S_{135} (ver figura 2.2):

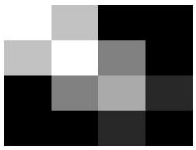
$$S_0 = \frac{1}{140}\begin{bmatrix} 40 & 10 & 0 & 0 \\ 10 & 32 & 12 & 0 \\ 0 & 12 & 16 & 4 \\ 0 & 0 & 4 & 0 \end{bmatrix}$$



$$S_{45} = \frac{1}{120}\begin{bmatrix} 4 & 26 & 0 & 0 \\ 26 & 8 & 20 & 0 \\ 0 & 20 & 8 & 4 \\ 0 & 0 & 4 & 0 \end{bmatrix}$$



$$S_{90} = \frac{1}{132}\begin{bmatrix} 24 & 18 & 0 & 0 \\ 18 & 24 & 12 & 0 \\ 0 & 12 & 16 & 4 \\ 0 & 0 & 4 & 0 \end{bmatrix}$$



$$S_{135} = \frac{1}{120}\begin{bmatrix} 4 & 26 & 0 & 0 \\ 26 & 8 & 20 & 0 \\ 0 & 20 & 8 & 4 \\ 0 & 0 & 4 & 0 \end{bmatrix}$$

Figura 2.2: Matrices de co-ocurrencia normalizadas.

2.1.1. Propiedades Estadísticas

Para la caracterización de textura, consideramos un conjunto de características derivadas de cuatro GLCM simétricas normalizadas: contraste (C), homogeneidad (H), media (M), energía (N) y varianza (V).

$$
\begin{aligned}
C &= \sum_{i=1}^{k}\sum_{j=1}^{k} S_{i,j}(i-j)^2 \\
H &= \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{S_{i,j}}{1+(i-j)^2} \\
M &= M_i = \sum_{i=1}^{k}\sum_{j=1}^{k} iS_{i,j} = M_j = \sum_{i=1}^{k}\sum_{j=1}^{k} jS_{i,j} \\
V &= \sum_{i=1}^{k}\sum_{j=1}^{k} S_{i,j}(i-M_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} S_{i,j}(j-M_j)^2 \\
N &= \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} S_{i,j}^2}
\end{aligned}
\tag{2.1}
$$

Donde $S_{i,j}$ representa el elemento (i,j) de una GLCM simétrica normalizada y k representa el numero de niveles de gris. Una clase de textura i consiste de un conjunto de n imágenes de miembros: $T_i = \{t_1, t_2, \cdots\cdots, t_n\}_i$. Para cada imagen miembro podemos calcular cuatro direccionales GLCM simétricas normalizadas:

$$\{(t_0^G, t_{45}^G, t_{90}^G, t_{135}^G)_1, (t_0^G, t_{45}^G, t_{90}^G, t_{135}^G)_2, \cdots, (t_0^G, t_{45}^G, t_{90}^G, t_{135}^G)_n\}_i$$

Podemos calcular características en la ecuación (2.1) para cada GLCM direccional. Cada característica se promedia sobre las cuatro GLCM direccionales para cada imagen miembro $\{(t_X^{-G})_1 \cdots (t_X^{-G})_n\}_i$ donde :

$$t_X^G = \frac{t_{X,0}^G + t_{X,45}^G + t_{X,90}^G + t_{X,135}^G}{4} \tag{2.2}$$

y $X \in \{C, H, M, V, N\}$.

Una clase de textura se respresenta por la colección de valores de sus características obtenidas durante una fase de entrenamiento. Una imagen de prueba s_j con sus características promedio calculado $(S_X^{-G})_j$, pertenece a una clase de textura específica si la probabilidad de que sus valores de características de ser un miembro de esa clase de entrenamiento es máxima.

2.2. Transformada Wavelet

2.2.1. Calculando una Matriz Discreta Wavelet (WDM)

Una wavelet es una función matemática utilizada para analizar una señal dependiente del tiempo a diferentes resoluciones. La transformada discreta wavelet (DWT) analiza la señal en diferentes resoluciones mediante la descomposición de la señal en un componente de aproximación y en un conjunto de componentes de detalle. La wavelet Haar [12] transforma una señal unidimensional x en un conjunto de promedios y diferencias

$$(x_1, x_2, \cdots, x_N) \rightarrow (s_1, \cdots, s_{\frac{N}{2}} | d_1, \cdots, d_{\frac{N}{2}}) \tag{2.3}$$

donde

$$s_K = \frac{x_{2K-1} + x_{2k}}{2}, d_K = \frac{x_{2K-1} - x_{2k}}{2}, k = 1, \cdots, \frac{N}{2}$$

Como ejemplo, la transformada wavelet Haar W_4 para una señal 1D de cuatro elementos $(x_1, x_2, x_3, x_4)^T$ es:

$$W_4 X = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_1 + x_2 \\ x_3 + x_4 \\ x_1 - x_2 \\ x_3 - x_4 \end{bmatrix} \tag{2.4}$$

Para una señal X en 2D ($N \times N$), la correspondiente transformada wavelet 2D W_N es:

$$W_N X W_N^T = \begin{bmatrix} A & V \\ H & D \end{bmatrix} \tag{2.5}$$

Aquí A es el componente de aproximación y H_p, V_p, y D_p son los componentes de detalle horizontal, vertical, y diagonal en el nivel p, respectivamente. Usamos la matriz A de un nivel específico como la matriz de datos para el siguiente nivel. La matriz de datos se particiona en celdas de 2×2 :

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

Y para cada celda, los componentes de aproximación y detalle se calculan como sigue:

$$A(i,j) = \frac{1}{4}(x_{11} + x_{12} + x_{21} + x_{22})$$
$$H(i,j) = \frac{1}{4}\{(x_{11} + x_{12}) - (x_{21} + x_{22})\}$$
$$V(i,j) = \frac{1}{4}\{(x_{11} + x_{21}) - (x_{12} + x_{22})\}$$
$$D(i,j) = \frac{1}{4}\{(x_{11} + x_{22}) - (x_{12} + x_{21})\}$$

Para una operación de descomposición wavelet, la colección de matrices A, H_p, V_p y $D_p(p = 1, 2, 3, \cdots)$ es la Matriz de descomposición Wavelet (WDM). A continuación se muestran los resultados obtenidos al aplicar la Transformada Wavelet de dos dimensiones:

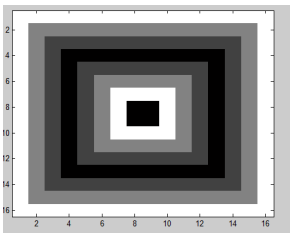


Figura 2.3: WDM Imagen original.

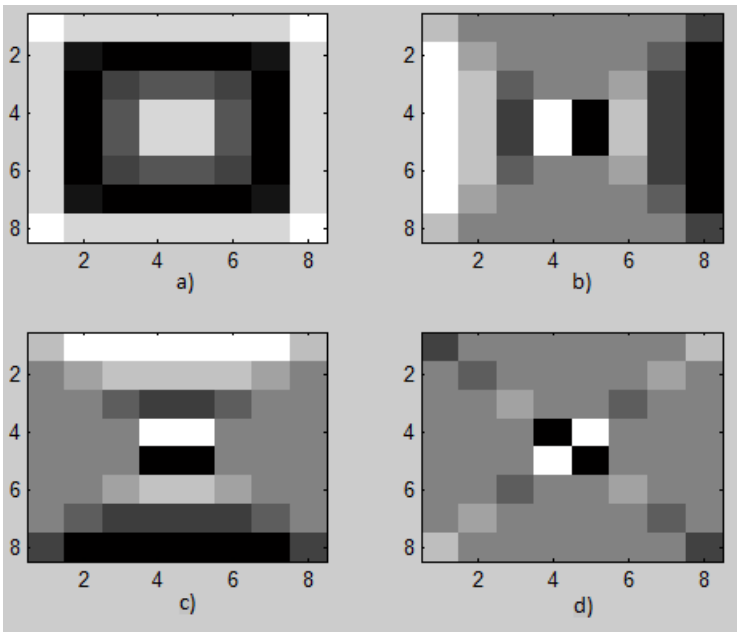


Figura 2.4: WDM Nivel 1 de descomposición: a) Componente de aproximación A, b) Componente detalle horizontal H, c) Componente detalle vertical V, d) Componente detalle diagonal D.

2.2.2. Características de clasificación basadas en la WDM

Una imagen de textura es descompuesta a p niveles usando una wavelet Haar 2D, produciendo la siguiente WDM:$A, H_1, V_1, D_1, \cdots Hp, Vp$ y D_p donde el subíndice indica el nivel.

A partir de los componentes de detalle, calculamos la matriz de covarianza:

$$C_{H_1}, C_{V_1}, C_{D_1}, \cdots, C_{H_p}, C_{V_p}, C_{D_p}$$

El elemento (i, j) ésimo de la matriz de covarianza $C_{i,j}$ para la matriz de datos

$$x = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \cdots & & & \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

Está definido en la ecuación 2.6

$$C_{i,j} = C_{j,i} = \frac{1}{m-1} \sum_{k=1}^{m} \{(x_{k,i} - \mu_i)(x_{k,j} - \mu_j)\} \tag{2.6}$$

donde

$$\mu_i = \frac{x_{1,i} + x_{2,i} + \cdots + x_{m,i}}{m} \text{ y } m \text{ es el número de filas en } X.$$

Un conjunto de matrices de correlación $R_{H_1}, R_{V_1}, R_{D_1}, \cdots, R_{H_p}, R_{V_p}, R_{D_p}$ se calcula a partir de la matriz de covarianza, donde el (i, j) ésimo elemento de una matriz de correlación es

$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} \cdot C_{j,j}}} \tag{2.7}$$

Los coeficientes wavelet combinados (Wp) de $(6p + 2)$ elementos que corresponden a la descomposicion del nivel p se calculan a partir de los componentes de aproximación, los componentes de detalle y las matrices de correlación, tal como se define en la ecuación 2.8. Este elemento se utiliza posteriormente como una característica para la discriminación de textura.

Así

$$\begin{aligned}\mathbf{W}_p = \{f_1, f_2, \cdot, f_{6p+2}\} = \{&\mu(A), \sigma(A), \mu(R_{H_1}), \sigma(H_1), \mu(R_{V_1}), \sigma(V_1), \mu(R_{D_1}), \sigma(D_1),\\ &\cdots\\ &\mu(R_{H_p}), \sigma(H_p), \mu(R_{V_p}), \sigma(V_p), \mu(R_{D_p}), \sigma(D_p)\}\end{aligned} \tag{2.8}$$

donde

$$\mu(x_1, x_2, \cdots, x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

y

$$\sigma(x_1, x_2, \cdots, x_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

La magnitud escalar de W_p se calcula a continuación:

$$W_p = |\mathbf{W}_p| = \sqrt{f_1^2 + f_2^2 + \cdots + f_{6p+2}^2} \tag{2.9}$$

Una clase de textura i consiste en un conjunto de n imágenes miembro:

$$T_i = \{t_1, t_2, \cdots, t_n\}_i \ .$$

Para cada imagen miembro, se calcula un coeficiente wavelet combinado mediante la ecuación 2.9 : $\{(t_w)_1, \cdots, (t_w)_n\}_i$

Una clase de textura está caracterizada por la colección de vectores T_i con valores obtenidos durante una fase de entrenamiento. Una imagen de prueba con sus características promedio calculadas $(S_W)_i$ pertenece a una clase de textura específica si la probabilidad de que el vector T_i sea miembro de la clase de entrenamiento es máxima.

Capítulo 3

Métodos de Clasificación de Células

Contenido

3.1. Introducción a la clasificación supervisada

Dentro del aprendizaje automático, encontramos técnicas de clasificación que nos permiten agrupar muestras de acuerdo a criterios o métodos, estas técnicas son la clasificación supervisada y la no supervisada.

El propósito fundamental, consiste en hacer una partición de un conjunto de objetos en categorías. Estas categorías, (o sus sinónimos: clases, conglomerados, grupos, etc.), se construyen de manera tal que un objeto en un grupo dado es similar, en algún sentido, a cualquier otro del mismo grupo; y objetos en distintos grupos tienden a ser diferentes.

Cada objeto es observado mediante un conjunto de variables cuantitativas que reflejan las cualidades fundamentales del mismo. Cada objeto tiene asociado entonces un conjunto de n valores sobre un conjunto de p variables, que en lo sucesivo se llamará una observación. El conjunto de observaciones se agrupa en una matriz X de dimensión $(n \times p)$.

Luego, el proceso de clasificar, que se lleva a cabo sobre la matriz X, consiste en: dado un conjunto de n observaciones y sus características dadas por p variables, se requiere agruparlos basándose en las semejanzas que existan entre sí.

3.1.1. Metodologías para abordar la clasificación

Las metodologías de clasificación provienen fundamentalmente de dos fuentes: el análisis estadístico multivariado y el área de la inteligencia artificial llamada computación emergente. Los métodos pueden organizarse así:

- Análisis estadístico multivariado
  - Análisis de conglomerados (cluster)
  - Análisis discriminante
- Computación emergente
  - Redes neuronales
    - Perceptrón multicapa
    - Mapas auto-organizativos
  - Lógica difusa

Gran parte de la teoría estadística del análisis multivariado, que constituye el núcleo de los procesos clasificatorios fue desarrollada en la primera mitad de este siglo. Sin embargo, dadas las dificultades de cálculo, sólo podían abordarse pequeños problemas: limitados tanto en el número de observaciones como en el de variables que caracterizaban a los objetos. Los algoritmos de computación emergente, que no exigen conocimiento previo del tipo de distribución de probabilidad, han probado ser muy eficientes para abordar problemas de datos complejo. En las últimas décadas los algoritmos de clasificación se implementan eficientemente sobre un computador y proveen los resultados sin intervención humana. Sin embargo, en la mayoría de las aplicaciones tecnológicas, el procesamiento obtenido es sólo un instrumento de soporte en la toma de decisiones.

3.1.2. Extractores de características

Si bien la capacidad de cálculo de los actuales computadores permite resolver eficientemente gran parte de los problemas de clasificación no es menos cierto que cada vez la complejidad de los problemas de clasificación va en aumento: tanto en el número n de observaciones a clasificar como en la dimensión p del espacio de variables que definen el objeto.

Problema de la dimensionalidad

La mayoría de los algoritmos clasificatorios padecen del síndrome de la dimensionalidad: probada eficiencia para problemas de dimensión reducida pero se vuelven ineficientes en problemas de gran escala. Es así que para espacios donde la dimensión p es excesiva se vuelve indispensable reducir la dimensionalidad del mismo. Los procedimientos que llevan a cabo esa función se denominan extractores de características.

Propósito

El objetivo fundamental de un extractor de características en procesos de clasificación es encontrar una transformación desde el espacio de dimensión p de las variables asociadas a cada observación en un espacio de dimensión inferior, denominado espacio de las características, que retenga de cada observación lo esencial de la información necesaria para el proceso de clasificación. Más precisamente: que el proceso clasificador de las observaciones en el espacio de la totalidad de las variables y en el espacio de las características conduzca a una división de las observaciones en las mismas clases o con diferencias insignificantes.

Obviamente, la terminología de espacio de las características obedece a que de las numerosas variables que representan la observación se extraen las características esenciales de las mismas.

Existen tres razones principales para aplicar un extractor de características:

La primera, la complejidad computacional de los algoritmos de clasificación se reduce sensiblemente al trabajar sobre un espacio de dimensión inferior.

La segunda, los métodos estadísticos de estimación se vuelven más confiables en un espacio de dimensión reducida.

La tercera, la posibilidad de que la dimensión del espacio de las características no exceda de tres, para permitir una visualización gráfica de las clases en juego.

3.1.3. Tipos de clasificación

Existe una división primaria en el concepto de clasificar:

- clasificación supervisada
- clasificación no supervisada.

La diferencia fundamental entre ambos métodos estriba en si se conoce o no la clase a la cual pertenece cada patrón (observación) de los datos.

Clasificación Supervisada

Este tipo de clasificación cuenta con un conocimiento a priori, es decir para la tarea de clasificar un objeto dentro de una categoría o clase contamos con modelos ya clasificados (objetos agrupados que tienen características comunes).

Características de la clasificación supervisada:

- El conjunto de datos está formado por tuplas atributo-valor
- El problema de clasificación puede ser biclásico (elegir entre dos clases) o multiclásico (entre muchas clases)
- Los atributos pueden ser continuos (valores enteros o reales) o discretos (etiquetas)
- Puede existir ruido (ejemplos mal clasificados)
- Pueden existir datos incompletos (missing values)

Podemos diferenciar dos fases dentro de este tipo de clasificación :

La primera fase consiste en el desarrollo o creación de una o varias reglas de decisión (diseño del clasificador): el conjunto cuyas clases ya están bien definidas se desglosa en un conjunto de entrenamiento o de aprendizaje (para el diseño del clasificador) y otro llamado de test o de validación (para clasificación), estos nos servirán para construir un modelo o regla general para la clasificación. Se diseña el clasificador con el conjunto de entrenamiento y se observa su capacidad para clasificar con el conjunto de validación.

En la segunda fase es el proceso en sí de clasificar los objetos o muestras de las que se desconoce la clase a las que pertenecen.

Ejemplos de clasificación supervisada son: el diagnóstico de enfermedades, predicción de quiebra o bancarrota en empresas, reconocimiento de caracteres escritos a mano, en la minería de datos, etc.

Tipos de clasificadores dentro de la clasificación supervisada:

- Vecinos más cercanos (K-NN)
- Máquinas de Vectores Soporte (SVM)
- Redes neuronales
- Redes bayesianas
- Árboles de decisión
- Reglas de clasificación
- Sistemas difusos

Criterios de evaluación de clasificadores:

- Matriz de confusión: muestra la distribución de los errores cometidos por un clasificador a lo largo de las distintas categorías del problema

		Verdadero		
		CLASE1	**CLASE2**	
Predicho	**CLASE1**	*a*	*c*	*p*CLASE1
	CLASE2	*b*	*d*	*p*CLASE2
		π CLASE1	π CLASE2	N

  donde:
  a: son los casos que pertenecen a la clase y el clasificador los definió en esa clase.
  b: son los casos que si pertenecen a la clase y el clasificador no los definió en esa clase.
  c: son los casos que no pertenecen a la clase pero el clasificador los definió en esa clase.
  d: son los casos que no pertenecen a la clase y el clasificador definió que no pertenecen a esa clase.

- Tasa de error:$(b + c)/N$
- Sensibilidad: $a/(a + c)$ proporción de verdaderos positivos
- Especificidad:$d/(b + d)$ proporción de verdaderos negativos

Criterios de validación de clasificadores:

- Holdout: Se divide el conjunto de casos en dos grupos: conjunto de entrenamiento (2/3) y conjunto de test (1/3). El conjunto de entrenamiento se usa para generar el clasificador y el de test para evaluarlo.
- Validación cruzada (cross-validation): Se divide el conjunto de casos en K subconjuntos del mismo tamaño. Se utilizan K-1 subconjuntos como datos de entrenamiento y 1 subconjunto como datos de test. Se repite para los K subconjuntos y se calcula la media de la evaluación. Suele utilizarse K=10.
- Dejar uno fuera (leave one out): validación cruzada con K igual al número de casos.
- Bootstrapping: el conjunto de entrenamiento se escoge como una muestra aleatoria con reemplazamiento.

Clasificación no supervisada

La clasificación es no supervisada cuando se dispone de un conjunto de objetos (observaciones), donde se desconoce tanto el número de clases en que es razonable particionarlo así como a qué clase pertenece cada observación. Este proceso de clasificación no supervisada, es significativamente más complejo que el de la supervisada ya que se desconocen las clases naturales, y dependerá de la habilidad para seleccionar.

3.2. Árbol de Clasificación y Regresión - CART

CART: Classification And Regression Trees

Breiman (1984) desarrolló el algoritmo CART cuyo resultado es, en general, un árbol de decisión, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos históricos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos.

CART es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos grupos hijos o nodos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado. Las divisiones se seleccionan de modo que "la impureza"de los hijos sea menor que la del grupo madre y éstas están definidas por un valor de una variable explicativa (Deconinck et al., 2006).

El objetivo es particionar la respuesta en grupos homogéneos y a la vez mantener el árbol razonablemente pequeño. Para dividir los datos se requiere un criterio de particionamiento el cual determinará la medida de impureza, ésta última establecerá el grado de homogeneidad entre los grupos.

El algoritmo CART lleva a cabo una búsqueda exhaustiva de todas las posibles cortaduras para minimizar el porcentaje de clasificación incorrecta.

Se pueden mencionar las siguientes ventajas de los árboles de clasificación y/o regresión:

a. Se obtiene conocimiento estructurado en forma de reglas de clasificación o de los valores de una variable de intervalo. Esto facilita interpretar en un lenguaje llano la caracterización de las clases o los valores de una variable de intervalo.

b. Al ser un procedimiento de análisis no paramétrico (distribution free procedure) no se requiere validar supuestos distribucionales de probabilidad.

c. Permite trabajar con todo tipo de variables predictoras: binarias, nominales, ordinales y de intervalo o razón.

d. Permite valores desconocidos para las variables predictoras en los individuos, tanto en la fase de construcción del árbol como en la de predicción.

e. En el caso de clasificación se puede establecer probabilidad a priori de las clases.

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos (Timofeev, 2004):

1. Construcción del árbol máximo.
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada ("cross-validation").

3.2.1. Construcción del árbol óptimo

El árbol máximo es construido utilizando un procedimiento de partición binario, comenzando en la raíz del árbol, este árbol es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es sobreajustado, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo.

Cada grupo es caracterizado por la distribución (respuesta categórica), o por la media (respuesta numérica) de la variable respuesta, el tamaño del grupo y los valores de las variables explicativas que lo definen. Gráficamente, el árbol se representa con el nodo raíz (los datos sin ninguna división) al iniciar y las ramas y hojas debajo (cada hoja es el final de un grupo). En la fase de construcción del árbol, se parte de la matriz de

$$\begin{bmatrix} x_{11} & \cdots & \cdots & \cdots & x_{1p} & y_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & \cdots & \cdots & x_{np} & y_n \end{bmatrix}$$

Figura 3.1: Ejemplo de una matriz de datos para construir un árbol.

datos (ver figura 3.1) donde n = número de observaciones y p = número de variables independientes, explicativas o predictoras.
El objetivo de este método será discriminar, estimar o predecir la variable Y en función de los predictores $X_1.....X_p$, mediante particiones sucesivas del conjunto de individuos, maximizando una medida de contenido de información respecto a la variable respuesta. En la fase de validación se puede utilizar esta misma matriz de diseño o entrenamiento u otra similar pero independiente (muestra de validación o prueba).

Un árbol es un conjunto de nodos y arcos. Cada nodo representa un subconjunto de la población. Distinguimos: Nodo raíz que representa a toda la población y no tiene arcos entrantes. Nodos terminales que representa la partición final. Nodos intermedios cuyos arcos salientes apuntan a los nodos hijos.

La presentación de la información se hace en un diagrama en forma de árbol invertido donde el proceso recursivo, muy esquemáticamente, se traduce en los siguientes pasos:

a. El nodo raíz es dividido en subgrupos (dos o más) determinados por la partición de una variable predictora elegida, generando nodos hijos.

b. Los nodos hijos son divididos usando la partición de una nueva variable. El proceso recursivo se repite para los nuevos nodos hijos sucesivamente hasta que se cumpla alguna condición de parada.

c. Algunos de los nodos resultantes son terminales, mientras que otros nodos continúan dividiéndose hasta llegar a un nodo terminal.

d. En cada árbol se cumple la propiedad de tener un camino único entre el nodo raíz y cada uno de los demás nodos del árbol.

Ejemplo: En 215 pacientes que sufrieron un ataque al corazón se evaluaron variables sociodemográficas, historia médica y exámenes de laboratorio. A los 30 días 37 pacientes murieron. Se presenta el árbol de clasificación desarrollado con el fin de estimar "El riesgo de un segundo ataque"(Fig.3.2).

En el proceso recursivo descrito se deben establecer algunos criterios:

a. Cómo son los cortes posibles y un número máximo de cortes determinados por un predictor desde el nodo. Los cortes que se establecen para variables ordinales y de intervalo se realizan por intervalos consecutivos.

b. Una condición de admisibilidad para los cortes posibles.

c. Una medida de contenido de información del árbol respecto al conjunto de individuos o un criterio de optimización de los cortes; es decir, obtener la mejor combinación de cortes admisibles respecto a una variable predictora.

d. Determinar la descripción de la variable objetivo en los nodos del árbol. Para clasificación: el grupo con la mayor representación determina la clase a la que asigna el nodo. En caso de empates se puede elegir cualquiera. Para regresión: en los nodos se estiman las medias muestrales de la variable respuesta condicionadas a los nodos.

e. Una condición de parada para un nodo de un árbol. Por ejemplo, si el número de individuos en el nodo es inferior a un valor pre-especificado, si la contribución del nodo a la calidad del árbol es mayor que otro umbral, si la profundidad del nodo es igual a un parámetro pre-especificado.

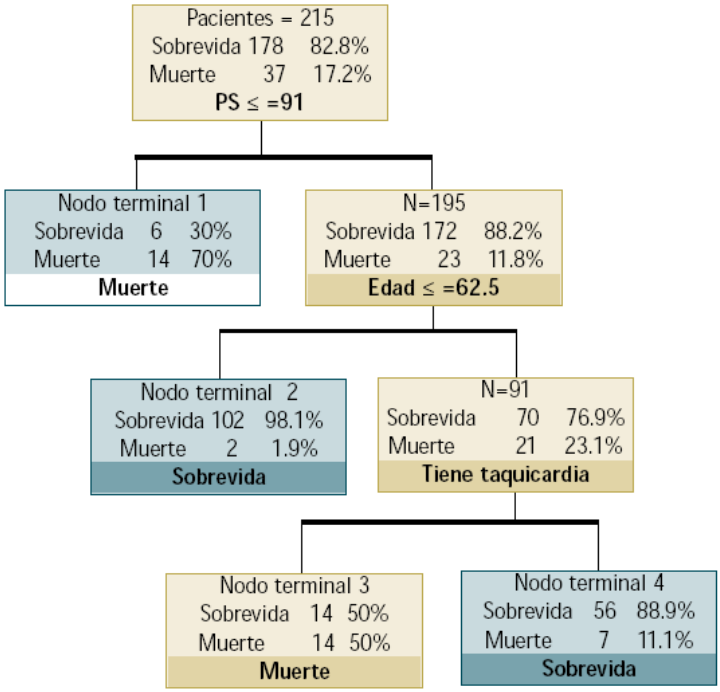


Figura 3.2: Ejemplo de árbol de clasificación.

El criterio más importante en la construcción del árbol es la elección de una medida de contenido de información del árbol con respecto a las clases o variable de intervalo de interés ya que la elección de este criterio diferencia los algoritmos de selección.

Medida de contenido de la información

Es la suma ponderada de una medida de contenido de la información de las hojas del árbol. Es una función de incertidumbre o entropía aplicada a una distribución de probabilidad. Al ser ésta una medida aditiva en los nodos, en un paso del algoritmo es suficiente con optimizar el incremento de la medida de contenido de información del árbol en el nodo que se está explorando. En este caso, se obtiene la combinación de cortes que hace máxima la reducción de la incertidumbre en los nodos del árbol. A continuacion se describen algunas medidas.

Calidad del Nodo: Función de Impureza

La función de impureza es una medida que permite determinar la calidad de un nodo, ésta será denotada por $i(t)$. Existen varias medidas de impureza (criterios de particionamiento) que nos permiten analizar varios tipos de respuesta, las tres medidas más comunes presentadas por Breiman et al. (1984), para árboles de clasificación son:

- El índice de información o entropía el cual se define como:

$$i(t) = \sum_{j} p(j|t)\mathrm{In}p(j|t) \tag{3.1}$$

El objetivo es encontrar la partición que maximice $\triangle i(t)$ en la ecuación 3.2

$$\triangle i(t) = \sum_{j=1}^{k} p(j|t)\mathrm{In}p(j|t) \tag{3.2}$$

donde $j = 1, ..., k$ es el número de clases de la variable respuesta categórica y $p(j|t)$ la probabilidad de clasificación correcta para la clase j en el nodo t.

- El índice Gini tiene la forma

$$i(t) = \sum_{j \neq j} p(j|t)p(j|t) \tag{3.3}$$

Encontrar la partición que maximice $\triangle i(t)$ en 3.4

$$\triangle i = -\sum_{j=1}^{k} [p_j(t)]^2 \tag{3.4}$$

Este índice es el más utilizado. En cada división el índice Gini tiende a separar la categoría más grande en un grupo aparte, mientras que el índice de información tiende a formar grupos con más de una categoría en las primeras decisiones.

- El índice "Towing ". A diferencia del índice Gini, Towing busca las dos clases que juntas formen más del 50 % de los datos, esto define dos "super categorías" en cada división para las cuales la impureza es definida por el índice Gini. Aunque el índice towing produce árboles más balanceados, este algoritmo trabaja más lento que la regla de Gini (Deconinck et al., 2006). Para usar el índice towing seleccione la partición s, que maximice

$$\frac{p_L p_R}{4}\left[\sum_j |p(j|t_L) - p(j|t_R)|\right]^2 \tag{3.5}$$

donde t_L y t_R representan los nodos hijos izquierdo y derecho respectivamente, p_L y p_R representan la proporción de observaciones en t que pasaron a t_L y a t_R en cada caso.

3.2.2. Poda del árbol

El árbol obtenido es generalmente sobreajustado por tanto es podado, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño "adecuado" del árbol. Breiman et al. (1984) introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol. Computacionalmente el procedimiento descrito es complejo. Una forma es buscar una serie de árboles anidados de tamaños decrecientes (De ath & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño. Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de costo complejidad , $R\sigma(T)$. Para cada árbol T, la función costo-complejidad se define como (Deconinck et ˜al., 2006):

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{3.6}$$

donde $R(T)$ es el promedio de la suma de los cuadrados entre los nodos,que puede ser la tasa de mala clasificación total o la suma de cuadrados de residuales total dependiendo del tipo de árbol, $\left|\tilde{T}\right|$ es la complejidad del árbol, definida como el número total de nodos del sub-árbol y α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero, Cuando $\alpha = 0$ se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol.
La función $R\alpha(T)$ siempre será minimizado por el árbol más grande, por tanto se necesitan mejores estimaciones del error, para esto Breiman et ˜al. (1984) proponen obtener estimadores "honestos " del error por "validación cruzada".

Computacionalmente el procedimiento es exigente pero viable, pues sólo es necesario considerar un árbol de cada tamaño, es decir, los árboles de la secuencia anidada.

3.2.3. Selección del árbol óptimo

De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad (De ath & Fabricius, 2000), por tanto se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada.

El objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones.

El procedimiento de validación cruzada puede implementarse de dos formas:

- Si se cuenta con suficientes datos se parte la muestra, sacando la mitad o menos de los datos y se construye la secuencia de árboles utilizando los datos que permanecen, luego predecir, para cada árbol, la respuesta de los datos que se sacaron al iniciar el proceso; obtener el error de las predicciones; seleccionar el árbol con el menor error de predicción.

En general no se cuenta con suficientes datos como para utilizar el procedimiento anterior, de modo que otra forma sería:

- Validación cruzada con partición en V. (v-fold cross validation, se menciona más adelante).

La idea básica de la "validación cruzada" es sacar de la muestra de aprendizaje una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto sacado es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como "datos nuevos". El desempeño entendido como el error de predicción, es acumulado para obtener el error medio absoluto del conjunto de prueba.

Como se mencionó anteriormente, para la metodología CART generalmente se utiliza validación cruzada con partición en V (v-fold cross validation), tomando $V = 10$ y el procedimiento es el siguiente:

- Dividir la muestra en diez grupos mutuamente excluyentes y de aproximadamente igual tamaño.
- Sacar un conjunto por vez y construir el árbol con los datos de los grupos restantes. El árbol es usado para predecir la respuesta del conjunto eliminado.
- Calcular el error estimado para cada subconjunto.
- Repetir los " ítems" dos y tres para cada tamaño de árbol.
- Seleccionar el árbol con la menor tasa de mala clasificación.

Al llegar a este punto se procede a analizar el árbol obtenido. La figura 3.2 es el diagrama de flujo del algoritmo CART.

Como ejemplo suponga el árbol y los datos de la Figura 3.3, donde se quiere determinar un conjunto de reglas que indiquen si un conductor vive o no en los suburbios.

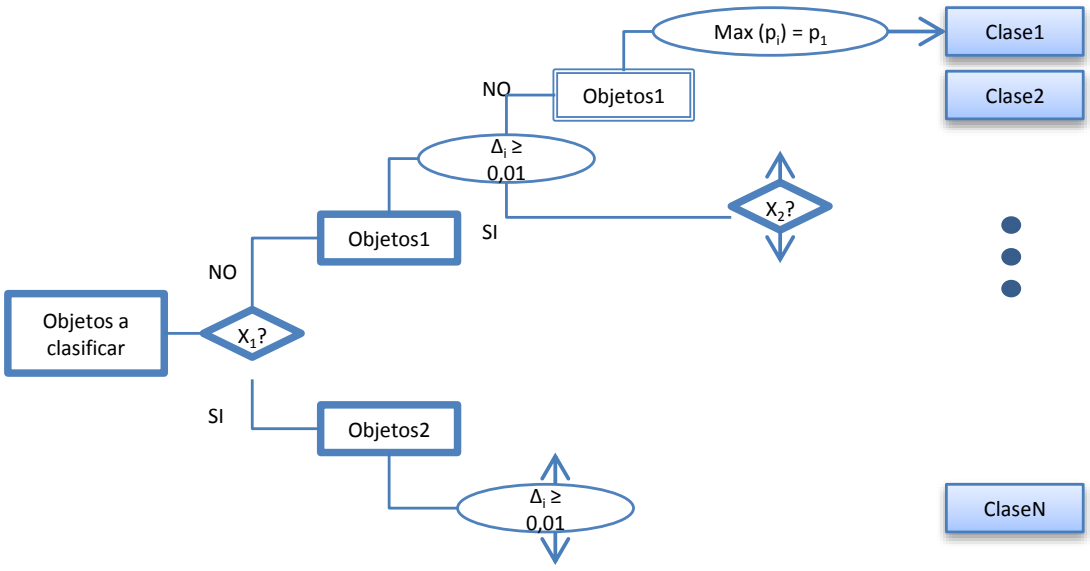
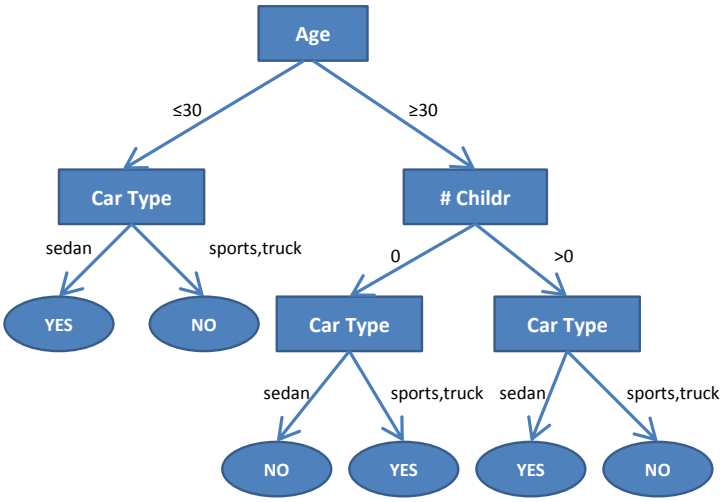


Figura 3.3: Diagrama de flujo del algoritmo CART.

Ejemplo árbol de clasificación. Fuente: Dobra (2002)
Del ejemplo de la figura 3.3 se concluye:



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

Figura 3.4: Ejemplo árbol de clasificación.

- Si Age $\leq$ 30 y CarType = Sedan entonces Si
- Si Age $\leq$ 30 y CarType =truck/Sports entonces No
- Si Age $>$ 30, Children = 0 y CarType = Sedan entonces No
- Si Age $>$ 30, Children = 0 y CarType =truck/Sports entonces Si
- Si Age $>$ 30, Children $>$ 0 y CarType = Sedan entonces Si
- Si Age $>$ 30, Children $>$ 0 y CarType =truck/Sports entonces No

Ventajas de los árboles de regresión:

- Las reglas de asignación son simples y legibles, por tanto la interpretación de resultados es directa e intuitiva.
- Es robusta frente a datos atípicos u observaciones mal etiquetadas.
- Es válida sea cual fuera la naturaleza de las variables explicativas: continuas, binarias nominales u ordinales.

- Es una técnica no parámetrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- Es computacionalmente rápido.

Desventajas de los árboles de regresión:

- Las reglas de asignación son muy sensibles a pequeñas perturbaciones en los datos (inestabilidad).
- Dificultad para elegir el árbol óptimo.
- Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.
- Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

3.3. Red Neuronal Probabilística

La RNP (sus siglas en ingles PNN - Probabilistic Neural Network) es una implementación de un algoritmo estadístico llamado núcleo de análisis discriminante en el que las operaciones se organizan en una red feedforward (alimentación hacia adelante) multicapa con cuatro capas.

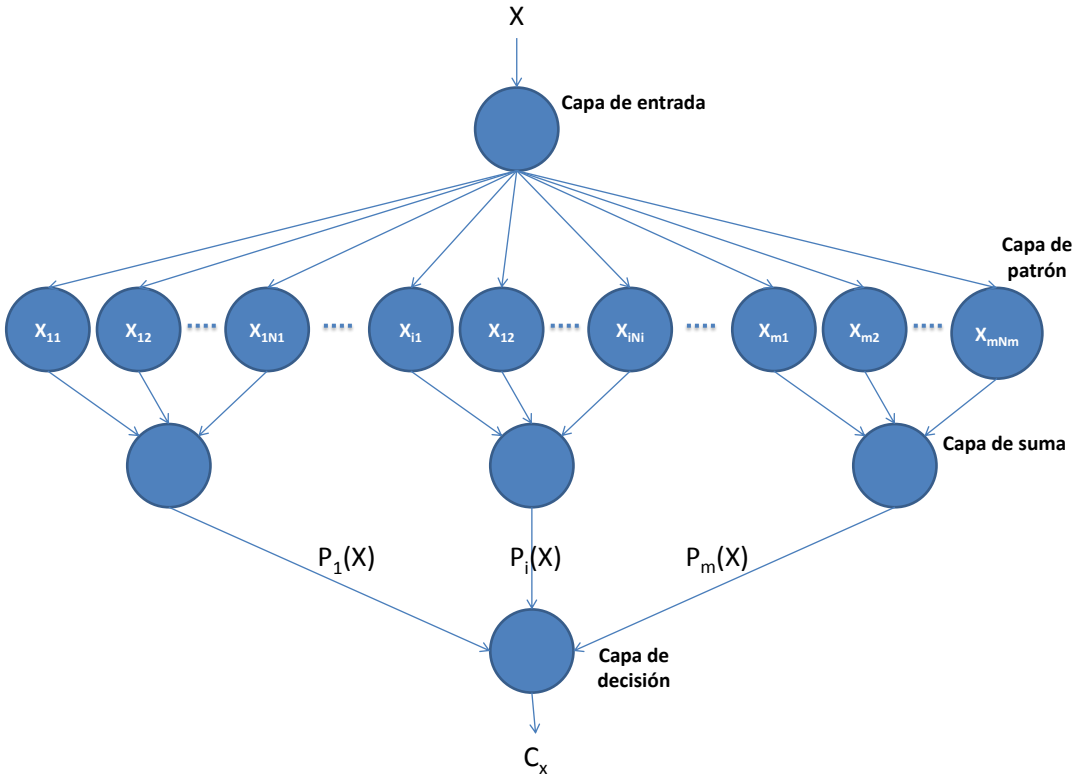


Figura 3.5: Arquitectura de una red neuronal probabilística.

La arquitectura de una RNP típica es como se muestra en la figura 3.4. La arquitectura de una RNP está compuesta de muchas unidades de procesamiento interconectadas o neuronas organizadas en capas sucesivas.

Está estrechamente relacionado con el estimador de la ventana parzen de función de densidad de probabilidad. Una PNN se compone de varias subredes, cada una de las cuales es una ventana estimador parzen de función de densidad de probabilidad para cada una de las clases, las cuatro capas se describen a continuación:

- Capa de entrada - Los nodos de entrada son el conjunto de mediciones.
- Capa del Patrón - Consiste de las funciones gaussianas formadas usando el conjunto dado de puntos de datos como centros.
- Capa de la Suma - Realiza una operación promedio de las salidas de la segunda capa para cada clase.
- Capa de Salida - Realiza una votación, seleccionando el valor más grande. Después se determina la etiqueta de la clase asociada.

La unidad de capa de entrada no realiza ningún cálculo y simplemente distribuye la entrada a las neuronas en la capa de patrón. A la recepción de un patrón de la capa de entrada, la neurona x_{ij} de la capa de patrón calcula su salida

$$\phi_{ij}(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}} \tag{3.7}$$

donde d denota la dimensión del vector de patrón x, σ es el parámetro de suavizado y x_{ij} es el vector de las neuronas. La suma de las neuronas de la capa calcula la probabilidad máxima de patrón x siendo clasificada en C_i resumiendo y promediando la salida de todas las neuronas que pertenecen a la misma clase

$$p_i(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \frac{1}{N_i} \sum_{j=1}^{N_i} e^{-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}} \tag{3.8}$$

donde N_i denota el número total de muestras en la clase C_i. Si las probabilidades a priori para cada clase son las mismas, y las pérdidas asociadas con la toma de una decisión incorrecta para cada clase son las mismas, la unidad de la capa de decisión clasifica el patrón x de acuerdo con la regla de decisión de Bayes, basada en la salida de la sumatoria de todas las capas de neuronas.

$$\widehat{C}(x) = \arg\max \{p_i(x)\}, i = 1, 2, \cdots, m \tag{3.9}$$

donde $\widehat{C}(x)$ denota la clase estimada del patrón de x y m es el número total de clases en las muestras de entrenamiento.

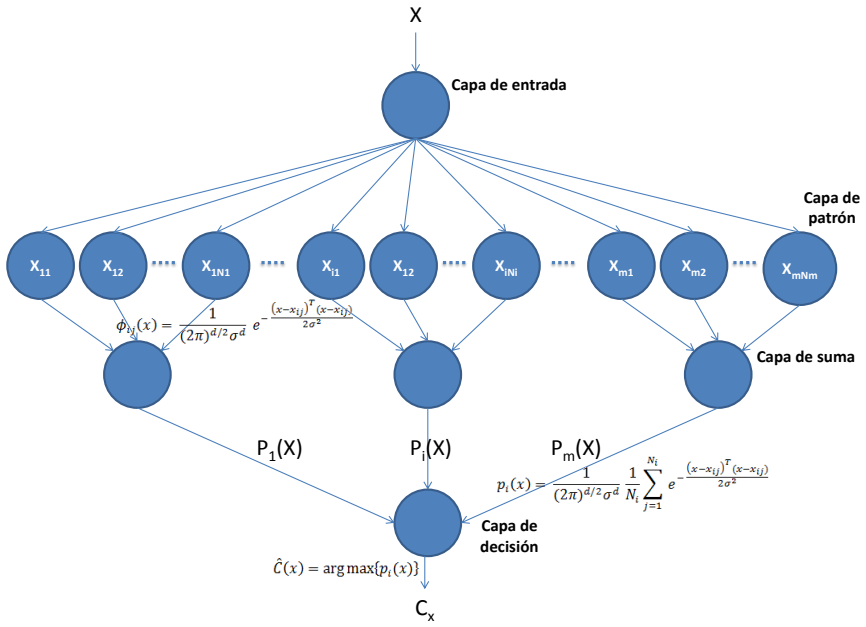


Figura 3.6: Representación de una red neuronal probabilística

VENTAJAS

- Rápido proceso de formación.
  - Órdenes de magnitud más rápidos que propagación hacia atrás.
- Una estructura inherentemente paralela.
- Garantizada para converger a un clasificador óptimo como el tamaño de los incrementos del conjunto representativo de entrenamiento.
  - Sin problemas locales mínimos
- Muestras de entrenamiento se pueden agregar o eliminar sin re-entrenamiento extenso.

DESVENTAJAS

- No es tan general como propagación hacia atrás.
- Grandes requisitos de memoria.
- Ejecución lenta de la red.
- Se requiere un conjunto de entrenamiento representativo.
  - Incluso más que otros tipos de NN.

3.4. Máquinas de Soporte Vectorial

3.4.1. Introducción a las Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial surgieron como un método de clasificación basado en la teoría de minimización del riesgo estructural de Vapnik. En la actualidad, tienen numerosas aplicaciones debido a su versatilidad y a sus prestaciones. Las SVM se han utilizado con éxito en campos como la recuperación de información, la categorización de textos, el reconocimiento de escritura o la clasificación de imágenes.

Para poder clasificar con las máquinas de soporte vectorial, se comienza realizando una etapa de aprendizaje. Consiste en encontrar el hiperplano $h(x) = 0$ que mejor separe un conjunto de datos $X \in R^D$ según la clase $Y \in \{-1, 1\}$ a la que pertenecen. Dicho hiperplano se corresponde con el que maximiza la distancia al punto más próximo de cada clase, por lo tanto, estará a la misma distancia de los ejemplos más cercanos entre ellos de cada categoría.

Según la teoría de Vapnik, el separador lineal que maximiza el margen (2 veces la distancia al punto más próximo de cada clase) es el que nos da la mayor capacidad de generalización, es decir, la capacidad de distinguir características comunes de los datos de cada clase que permitan clasificar imágenes que no sean las del conjunto de entrenamiento. Para hallarlo, es necesario resolver un problema de optimización usando técnicas de programación cuadrática.

A los datos que se utilizan para hallar la frontera de decisión (el hiperplano), se les conoce como vectores de entrenamiento o de aprendizaje.

A partir de unos datos de entrada x_i, las SVM nos proporcionarán su clase según la regla de clasificación $f(x_i) = \text{signo } (h(x_i))$.

Tras la fase de aprendizaje, se comprueba el error cometido tomando otra muestra de datos (denominados conjunto de test o validación) y comparando la salida que obtenemos con su clase real.

3.4.2. Clasificación binaria linealmente separable

Tenemos L puntos de entrenamiento, donde cada entrada x_i tiene D atributos (es decir su dimensionalidad es D) y están en la clases $y_i = -1$ ó $+1$.
Es decir nuestros datos de entrenamiento son de la forma:

$$\{x_i, y_i\} \text{ donde } i = 1 \cdots L, y_i \in \{-1, 1\}, x_i \in R^D$$

Aquí asumimos que el dato es linealmente separable, significa que podemos dibujar una línea en una gráfica de x_1 contra x_2 separando las dos clases cuando $D = 2$ y un hiperplano en gráficas de $x_1, x_2, \cdots, x_D$ para cuando $D > 2$.

Este hiperplano puede ser descrito por $w \cdot x + b = 0$ donde:

- w es normal al hiperplano
- $\frac{b}{\|w\|}$ es la distancia perpendicular desde el hiperplano hasta el origen

Los vectores de soporte son ejemplos cercanos a la separación del hiperplano y el objetivo de las maquinas de soporte vectorial (SVM) es orientar este hiperplano de tal manera que esté lo más alejado posible de los miembros más cercanos que pertenecen a distintas clases.

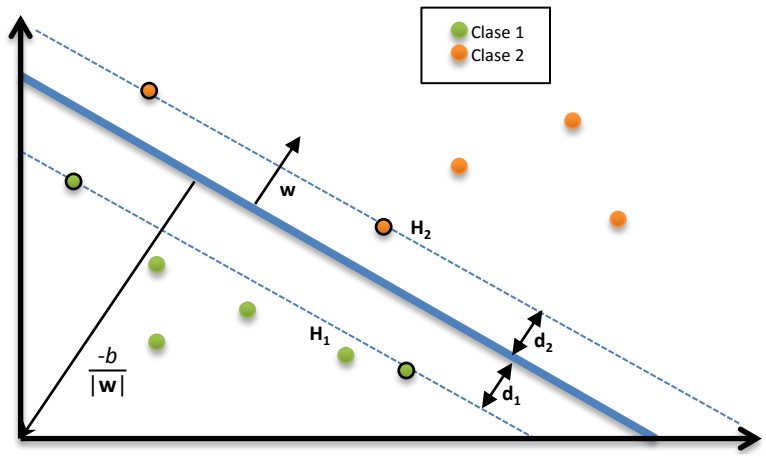


Figura 3.7: Representación de hiperplano entre dos clases linealmente separables

Refiriéndose a la Figura 3.5, implementar una SVM significa obtener las variables w y b tales que los datos de entrenamiento se pueden describir:

$$x_i \cdot w + b \geq +1 \qquad \text{para} \qquad y_i = +1 \tag{3.10}$$

$$x_i \cdot w + b \geq -1 \qquad \text{para} \qquad y_i = -1 \tag{3.11}$$

Estas ecuaciones se pueden combinar para obtener:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall_i \tag{3.12}$$

Si sólo consideramos los puntos que se encuentran más cerca al hiperplano de separación, es decir los vectores de soporte (mostrados en círculos en el diagrama), entonces los dos planos H_1 y H_2 en los que caen estos puntos se pueden describir por:

$$x_i \cdot w + b = +1 \qquad \text{para } H_1 \tag{3.13}$$

$$x_i \cdot w + b = -1 \qquad \text{para } H_2 \tag{3.14}$$

Además, podemos definir d_1 como la distancia desde H_1 al hiperplano y d_2 desde H_2 a éste. La equidistancia del hiperplano desde H_1 y H_2 significa que $d_1 = d_2$, una cantidad conocida como margen de SVM.

Para orientar el hiperplano de tal forma que esté tan lejos como sea posible de los vectores de soporte, necesitamos maximizar este margen.

La geometría simple de vector muestra que el margen es igual a $\frac{1}{\|w\|}$ y su maximización está sujeta a la restricción en (3.12), que es equivalente a encontrar:

$$min\, \|w\| \text{ de tal manera que } y_i(x_i \cdot w + b) - 1 \geq 0\ \forall_i$$

Minimizar $\|w\|$ es equivalente a minimizar $\frac{1}{2}\|w\|^2$ y el uso de este término hace posible resolver el problema utilizando la Programación Cuadrática (QP). Por lo tanto necesitamos encontrar:

$$\min \frac{1}{2}\|w\|^2 \qquad \text{sujeto a} \qquad y_i(x_i \cdot w + b) - 1 \geq 0 \qquad \forall_i \tag{3.15}$$

Con el fin de atender a las restricciones en esta minimización, tenemos que asignar multiplicadores LaGrange α donde $\alpha_i \geq 0 \quad \forall_i$:

$$\begin{aligned} \text{Lp} &\equiv \frac{1}{2}\|w\|^2 - \alpha\left[y_i(x_i \cdot w + b) - 1\forall_i\right] && (3.16)\\ &\equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L}\alpha_i\left[y_i(x_i \cdot w + b) - 1\right] && (3.17)\\ &\equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L}\alpha_i y_i(x_i \cdot w + b) + \sum_{i=1}^{L}\alpha_i && (3.18) \end{aligned}$$

Queremos encontrar w y b que minimizan, y α que maximiza (3.18) (manteniendo $\alpha_i \geq 0 \quad \forall_i$). Podemos hacer esto diferenciando L_P con respecto a w y b y haciendo las derivadas igual a cero:

$$\frac{\partial L_P}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{L}\alpha_i y_i x_i \tag{3.19}$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L}\alpha_i y_i = 0 \tag{3.20}$$

Substituyendo (3.19) y (3.20) en (3.18) obtenemos una nueva formulación que, al ser dependiente de α, debemos maximizar:

$$\begin{aligned} L_D &\equiv \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i \cdot x_j \text{ s.t. } \alpha_i \geq 0 \quad \forall_i, \sum_{i=1}^{L}\alpha_i y_i = 0 && (3.21)\\ &\equiv \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i H_{i,j}\alpha_j \text{ donde } H_i \equiv y_i y_j x_i \cdot x_j && (3.22)\\ &\equiv \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\alpha^T H\alpha \text{ s.t. } \alpha_i \geq 0 \quad \forall_i, \quad \sum_{i=1}^{L}\alpha_i y_i = 0 && (3.23) \end{aligned}$$

Esta nueva formulación L_D, se conoce como la forma dual de la L_P primario. Vale la pena notar que calcular la forma dual requiere sólo el producto escalar de cada vector de entrada x_i, esto es importante para el "truco del núcleo" (kernel trick)

Después de haber pasado de minimizar L_P para maximizar L_D, necesitamos encontrar:

$$\max \alpha = \left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha\right] \quad \text{s.t.} \quad \alpha_i \geq 0 \quad \forall_i \quad \text{y} \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.24}$$

Este es un problema de optimización convexa cuadrática, y se utiliza un método QP para resolver α y de (3.19) nos dará w. Lo que queda es calcular b.

Cualquier dato satisface (3.20), que es un x_s vector de soporte, tendrá la forma:

$$y_s(x_s \cdot w + b) = 1$$

Substituyendo en (3.19):

$$y_s(\sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b) = 1$$

Donde S denota el conjunto de índices de los vectores de soporte. S es determinado por encontrar los índices i donde $\alpha_i > 0$. Multiplicando por y_s y entonces utilizando $y_s^2 = 1$ de (3.10) y (3.11):

$$y_s^2(\sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b) = y_s$$

$$b = y_s - \sum_{m \in S} \alpha_m y_m x_m \cdot x_s$$

En lugar de utilizar un vector de soporte x_s arbitrario, es mejor tomar un promedio sobre todos los vectores de soporte en S:

$$b = \frac{1}{N_s} \sum_{m \in S} (y_s - \sum_{m \in S} \alpha_m y_m x_m \cdot x_s) \tag{3.25}$$

Así tenemos las variables w y b que definen nuestra orientación óptima de hiperplano de separación y por lo tanto, nuestra máquina de vectores de soporte.

3.4.3. Máquinas de soporte vectorial para clasificación multiclase

Las SVM se usan habitualmente para problemas de tipo binario. Una de las soluciones para resolver este problema multiclase es convertirlo en varios binarios. Para ello, existen 2 métodos distintos:

- Clasificación 1-v-r (del inglés one-versus-rest): en cada uno de los problemas se considera una clase positiva y las demás negativas, por lo que habrá que hallar tantos hiperplanos como clases existan. Es decir una clase versus todas las clases, por ejemplo para 3 clases se requiere crear tres SVM así:
  - SVM 1 vs. 2+3
  - SVM 2 vs. 1+3
  - SVM 3 vs. 1+2
- Clasificación 1-v-1 (del inglés one-versus-one): para cada problema se toman 2 clases de las K totales. Se compara cada clase con cada una de las restantes, lo que supone realizar $K(K-1)/2$ clasificaciones. Es decir una clase versus otra clase, por ejemplo con tres clases:
  - SVM 1 vs. 2
  - SVM 1 vs. 3
  - SVM 2 vs. 3

Capítulo 4

Resultados

Contenido

En particular clasificaremos los tipos de células en seis clases distintas que son : Centrómero, Citoplasmático, Homogénea, Moteada Gruesa, Moteada Fina y Nucleolar, como se muestra en la figura 4.1.

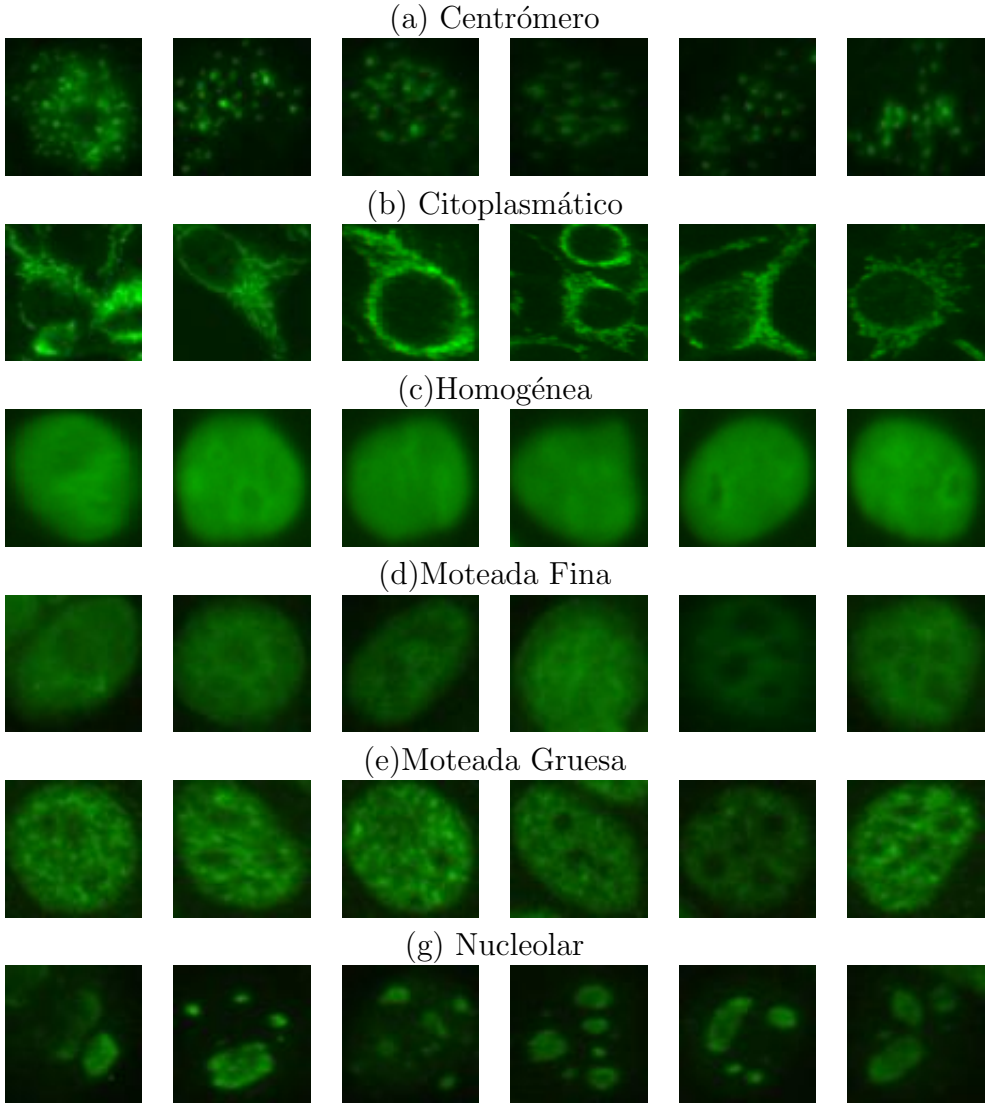


Figura 4.1: Clases de células.

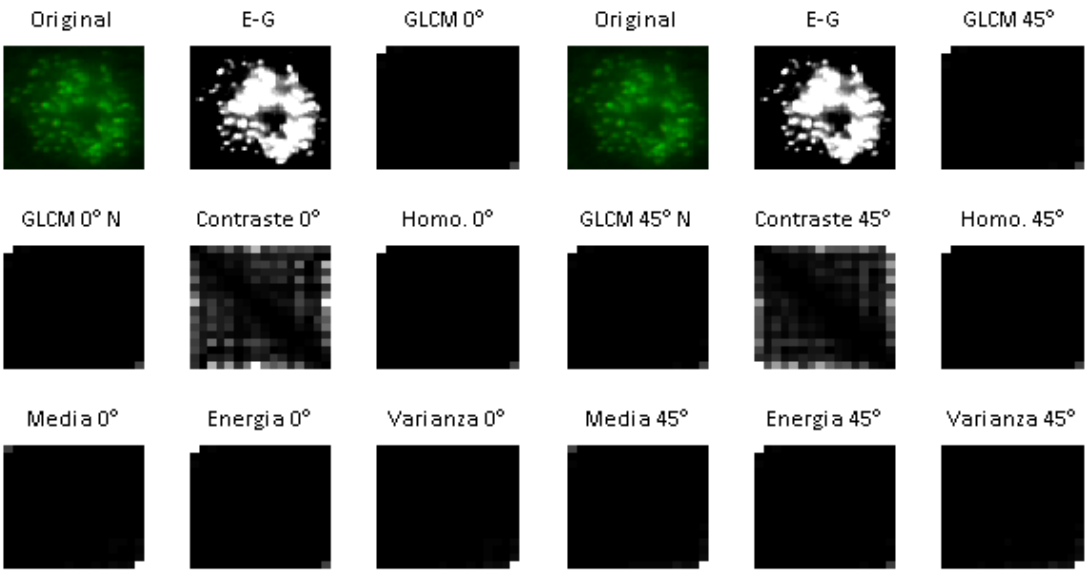
Una vez definidas las clases que serán utilizadas, se extraen las características de textura utilizando la Matriz de Coocurrencia de Niveles de Gris (Gray Level Coocurrence Matrix - GLCM).

Posteriormente se extraen características usando la Matriz Wavelet Discreta (Wavelet Discrete Matrix - WDM) de dos dimensiones, para ejemplificar se muestran los resultados de manera visual de dos células por clase (una en donde se distingue mejor la textura y otra en la que no).

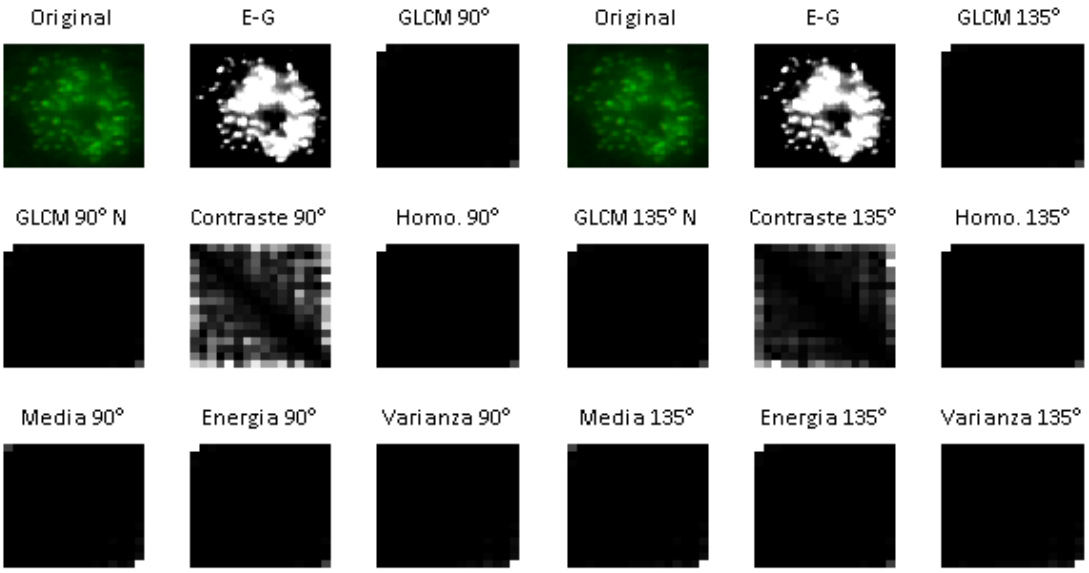
A continuación dos ejemplos por cada clase de células, primero de las GLCM en donde se muestra las cuatro direccionales y sus respectivas características, después de la WDM en todos sus niveles de descomposición.

4.1. Matriz de Coocurrencia de Niveles de Gris

CÉLULA CENTRÓMERO (imagen favorable)



GLCM 0 °Centrómero GLCM 45 °Centrómero

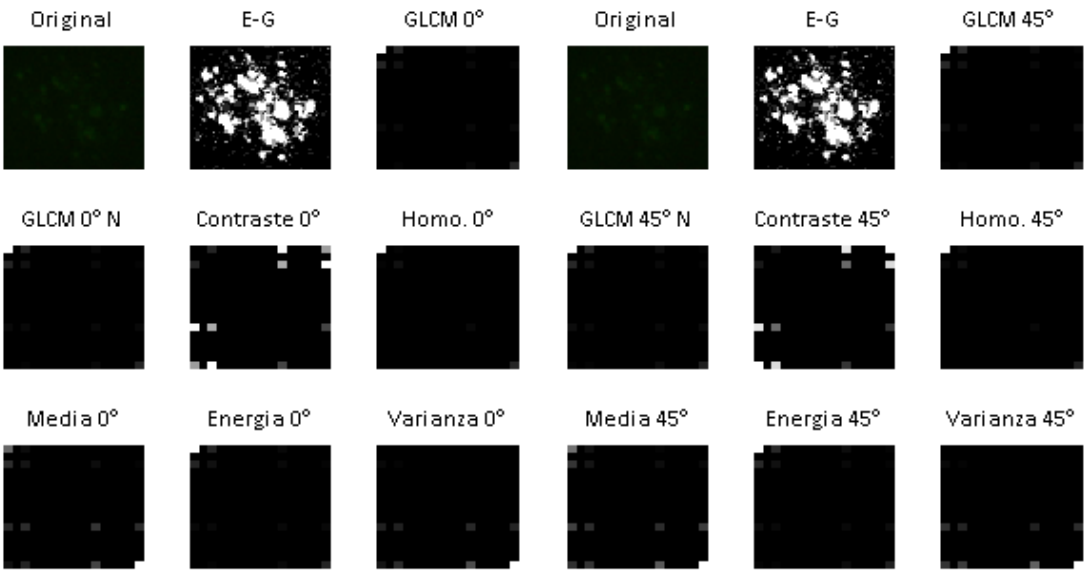


GLCM 90 °Centrómero GLCM 135°Centrómero

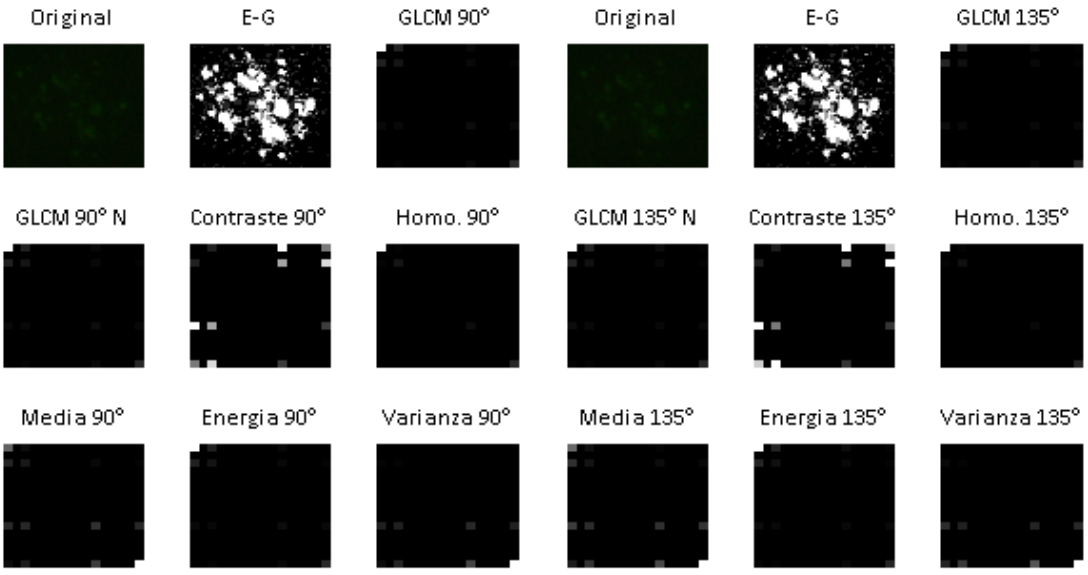
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.2: GLCM centrómero (favorable) .

CÉLULA CENTRÓMERO (imagen desfavorable)



GLCM 0 °Centrómero GLCM 45 °Centrómero

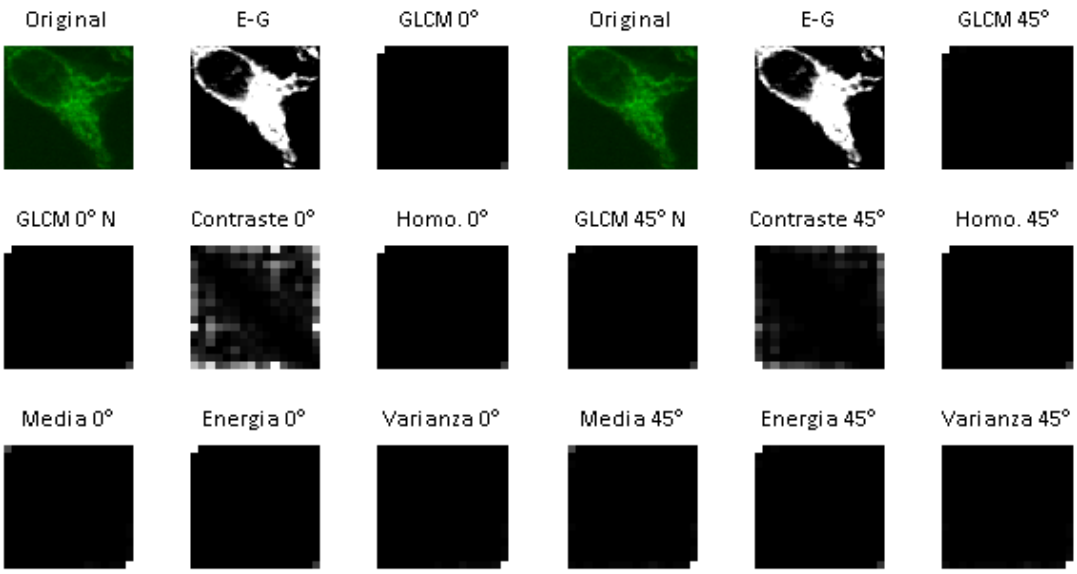


GLCM 90 °Centrómero GLCM 135°Centrómero

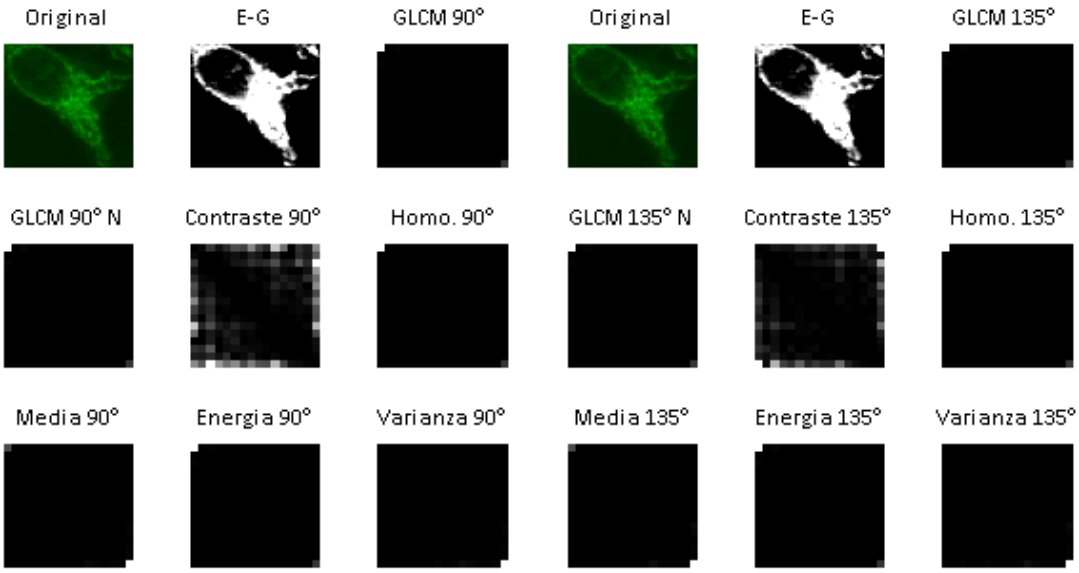
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.3: GLCM centrómero (desfavorable) .

CÉLULA CITOPLASMÁTICO (imagen favorable)



GLCM 0 °Citoplasmático GLCM 45 °Citoplasmático

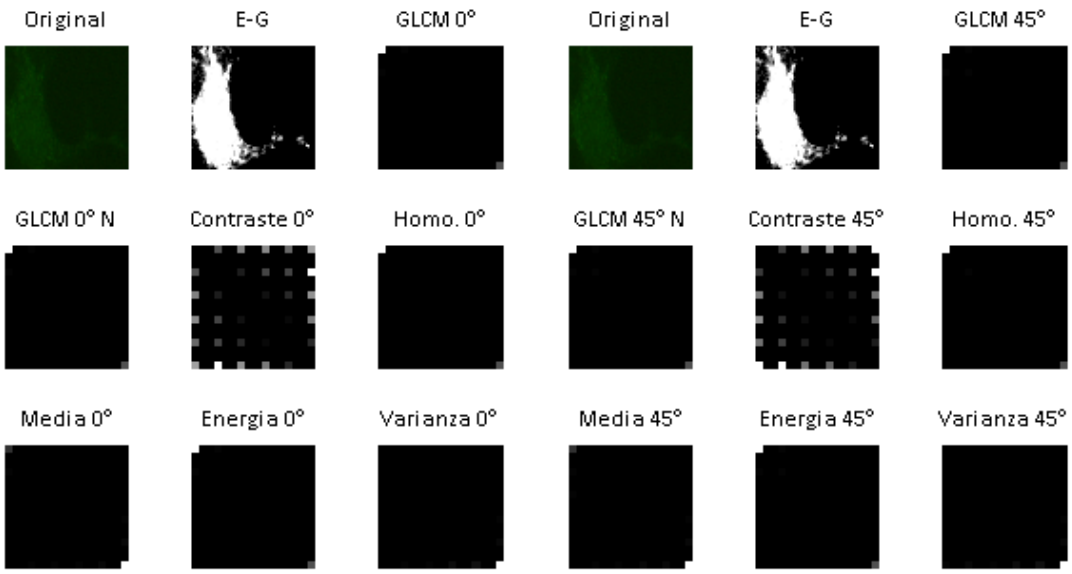


GLCM 90 °Citoplasmático GLCM 135°Citoplasmático

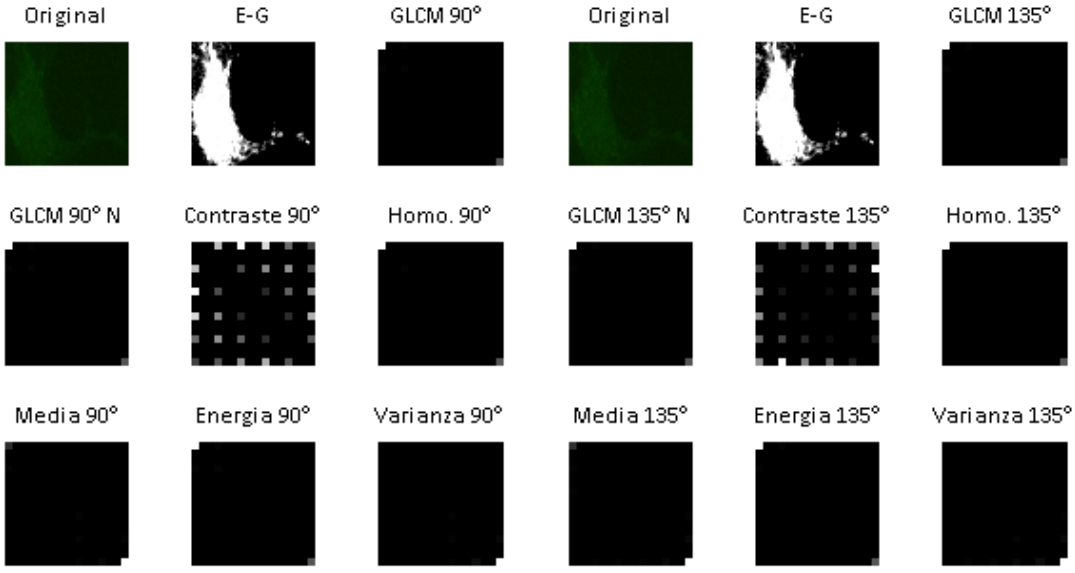
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.4: GLCM citoplasmático (favorable) .

CÉLULA CITOPLASMÁTICO (imagen desfavorable)



GLCM 0 °Citoplasmático GLCM 45 °Citoplasmático

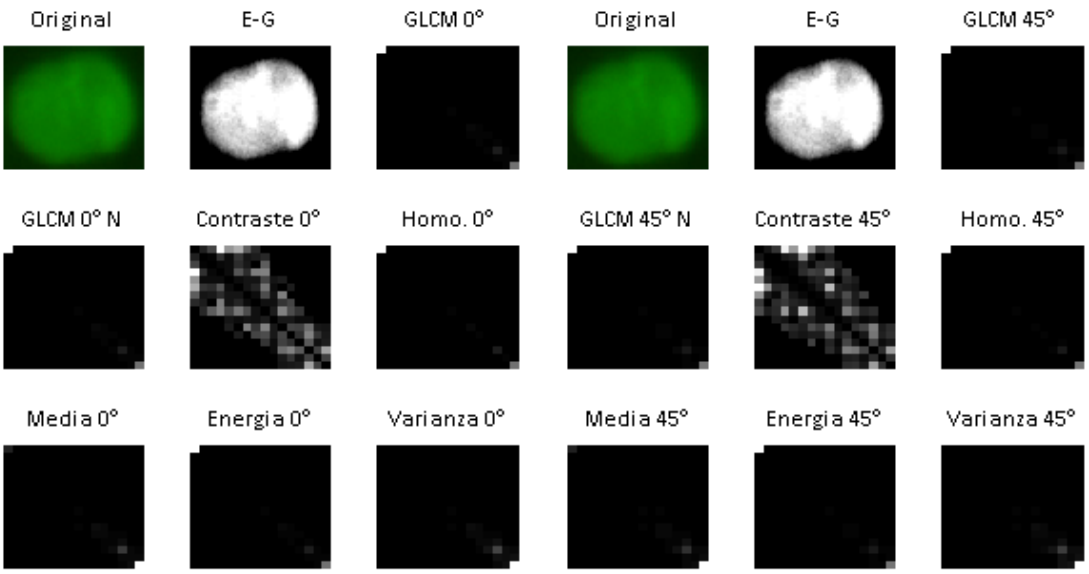


GLCM 90 °Citoplasmático GLCM 135°Citoplasmático

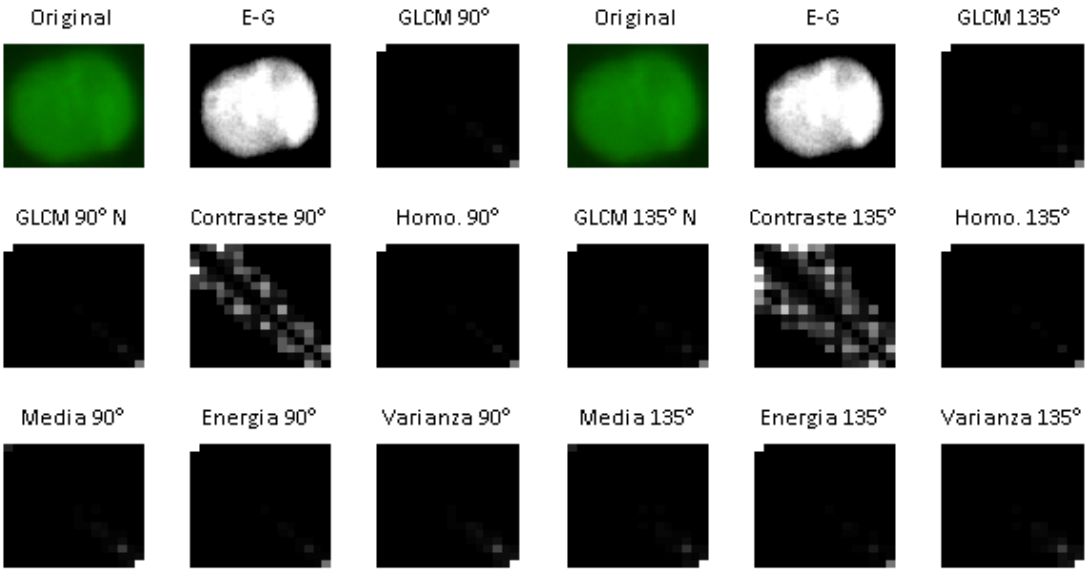
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.5: GLCM citoplasmático (desfavorable) .

CÉLULA HOMOGÉNEO (imagen favorable)



GLCM 0 °Homogéneo GLCM 45 °Homogéneo

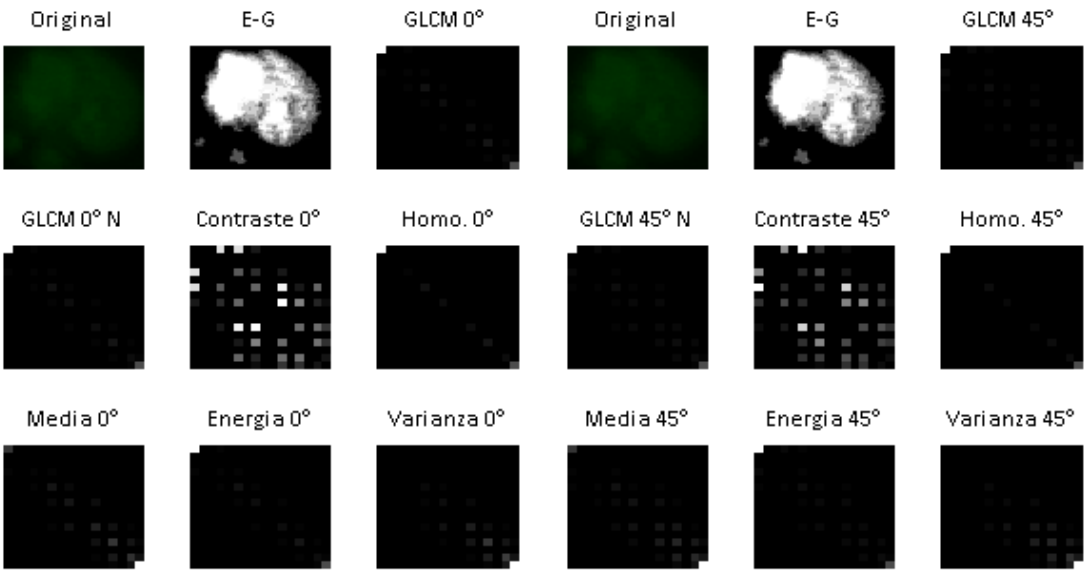


GLCM 90 °Homogéneo GLCM 135°Homogéneo

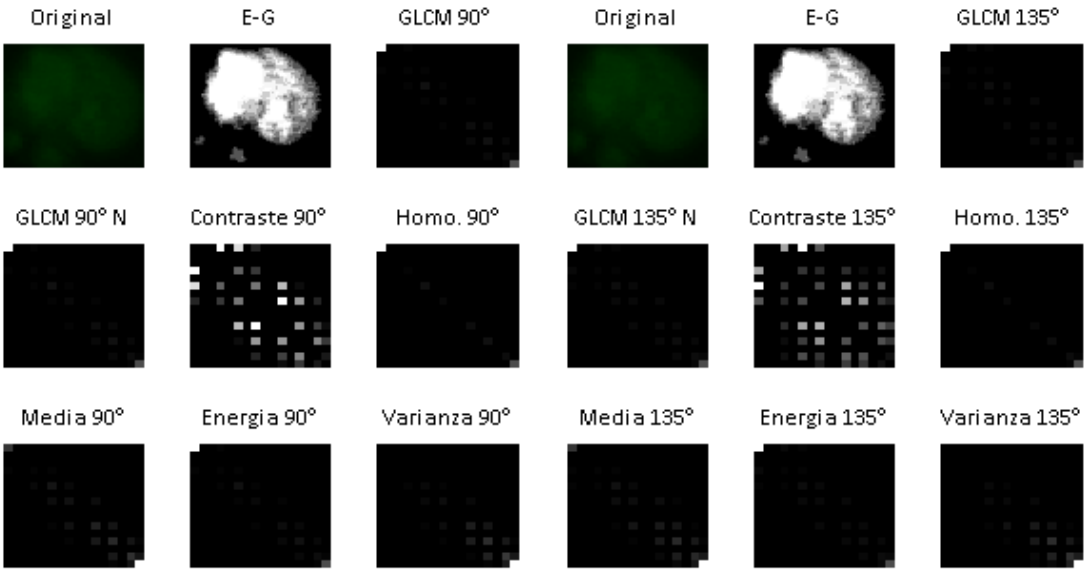
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.6: GLCM homogéneo (favorable) .

CÉLULA HOMOGÉNEO (imagen desfavorable)



GLCM 0 °Homogéneo GLCM 45 °Homogéneo

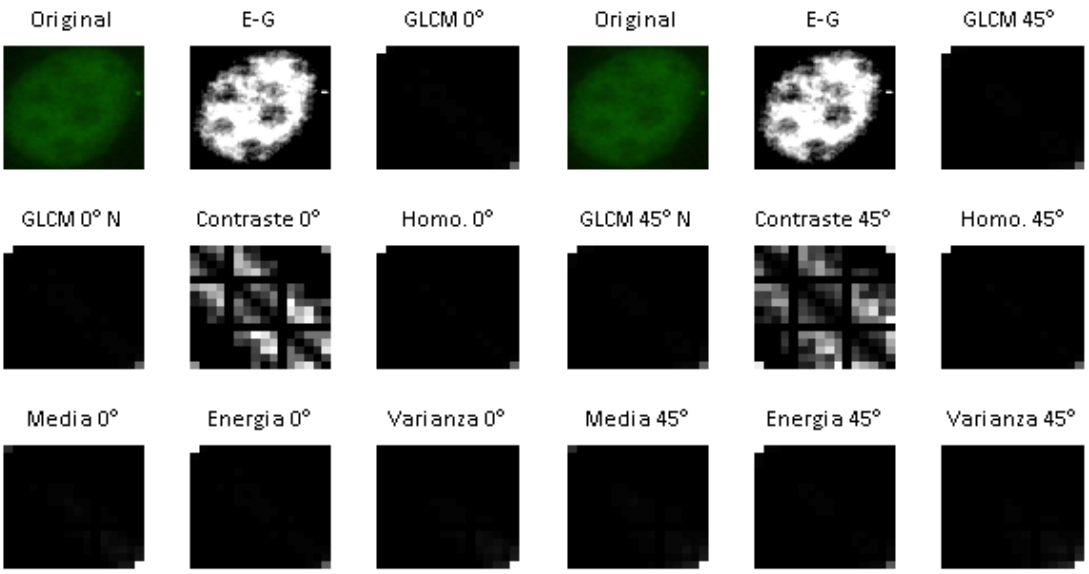


GLCM 90 °Homogéneo GLCM 135°Homogéneo

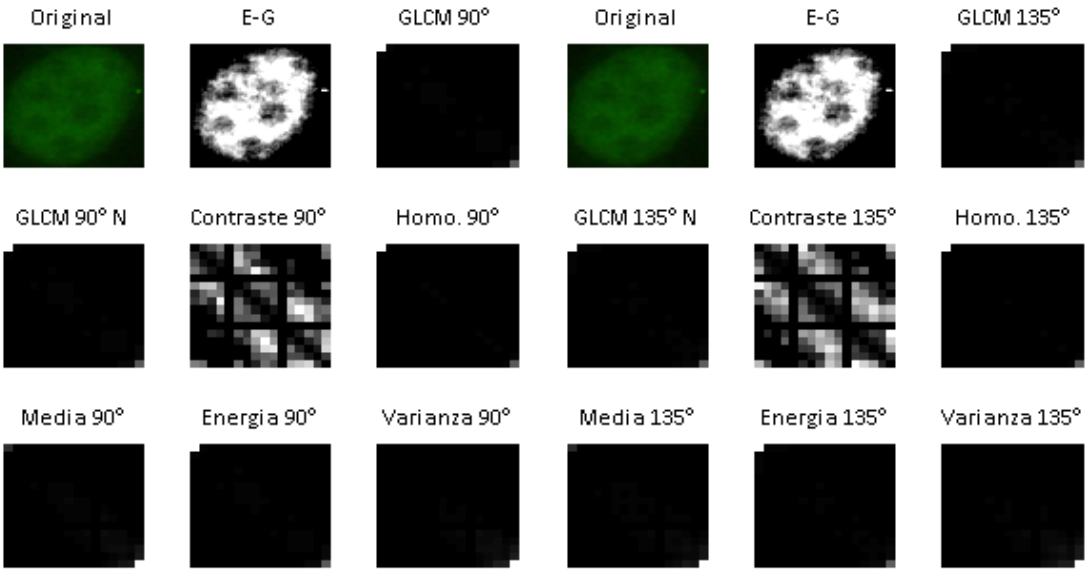
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.7: GLCM homogéneo (desfavorable) .

CÉLULA MOTEADA FINA (imagen favorable)



GLCM 0 °Moteada fina GLCM 45 °Moteada fina

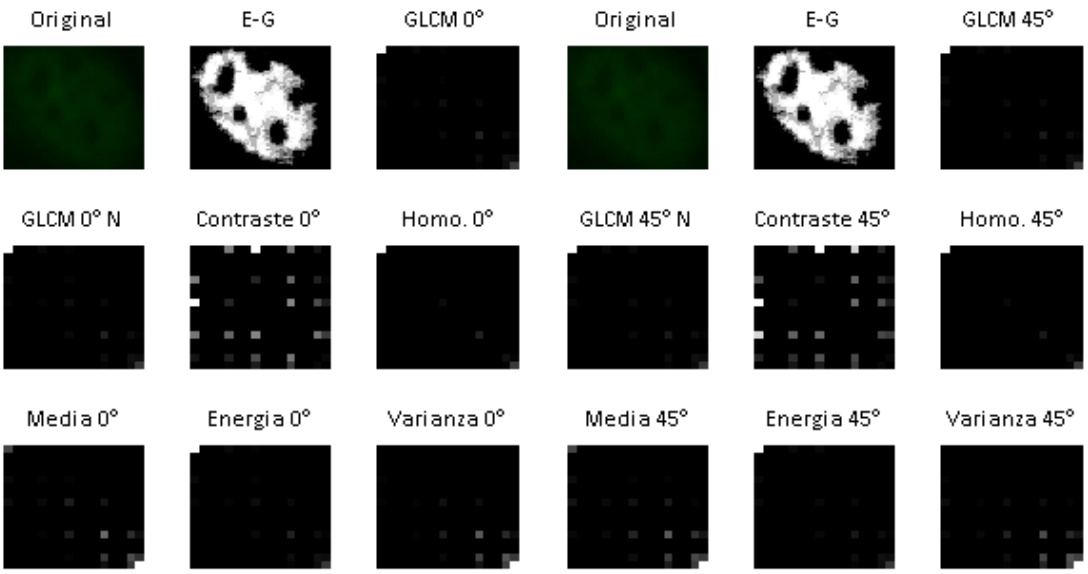


GLCM 90 °Moteada fina GLCM 135°Moteada fina

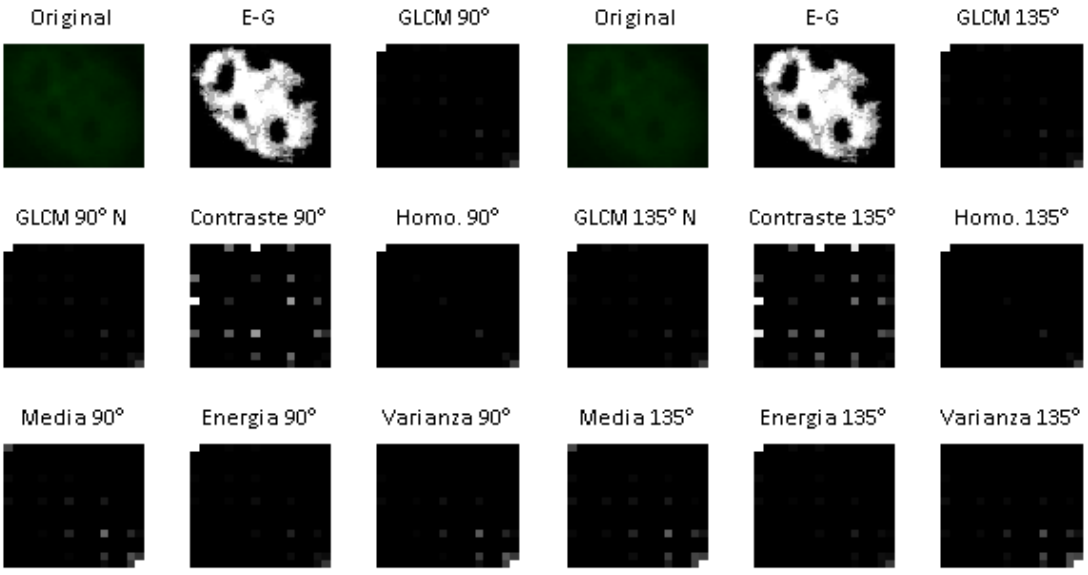
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.8: GLCM moteada fina (favorable) .

CÉLULA MOTEADA FINA (imagen desfavorable)



GLCM 0 °Moteada fina GLCM 45 °Moteada fina

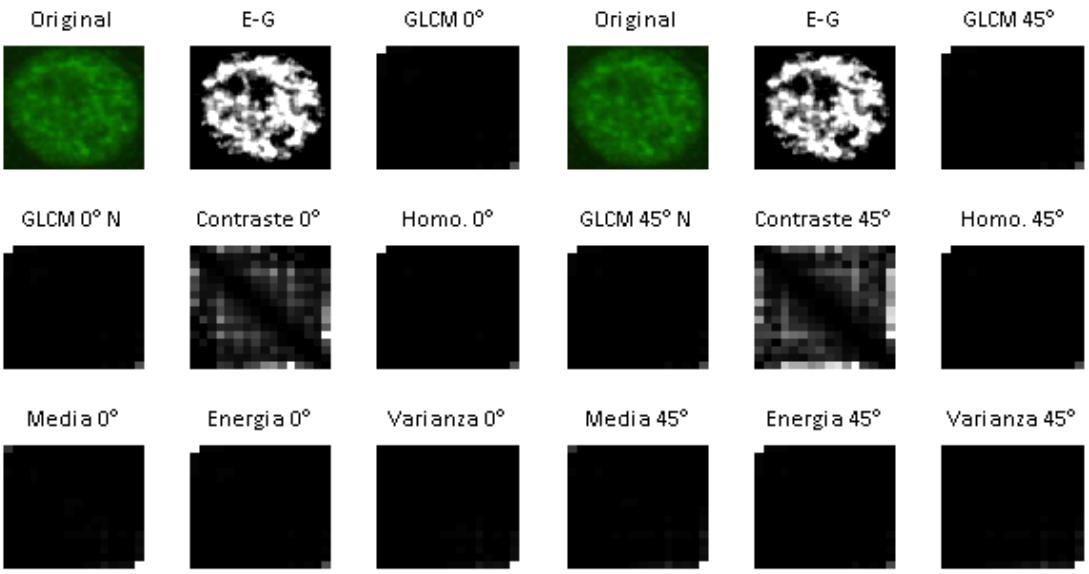


GLCM 90 °Moteada fina GLCM 135°Moteada fina

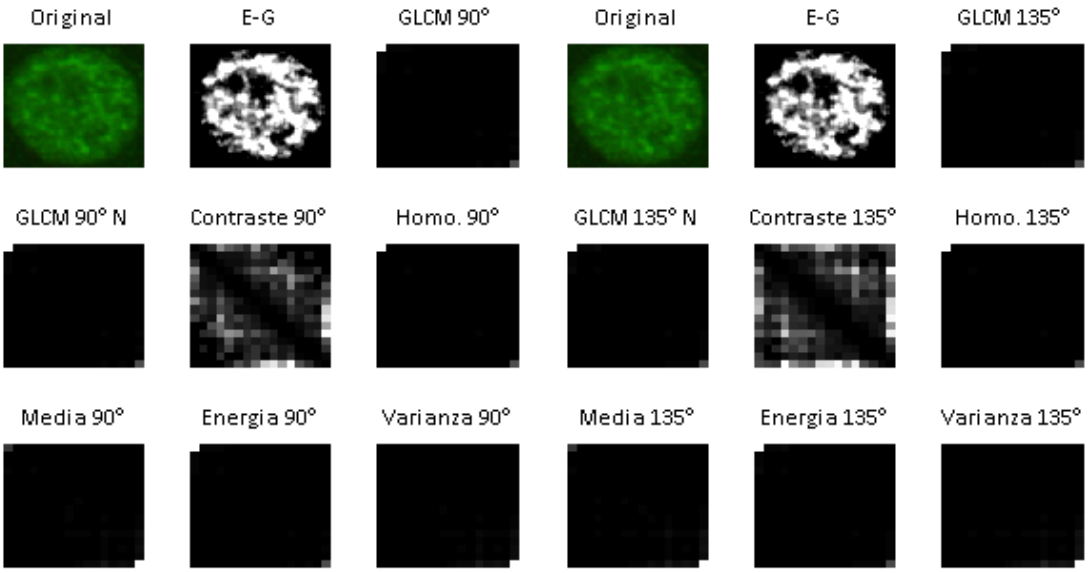
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.9: GLCM moteada fina (desfavorable) .

CÉLULA MOTEADA GRUESA (imagen favorable)



GLCM 0 °Moteada gruesa GLCM 45 °Moteada gruesa

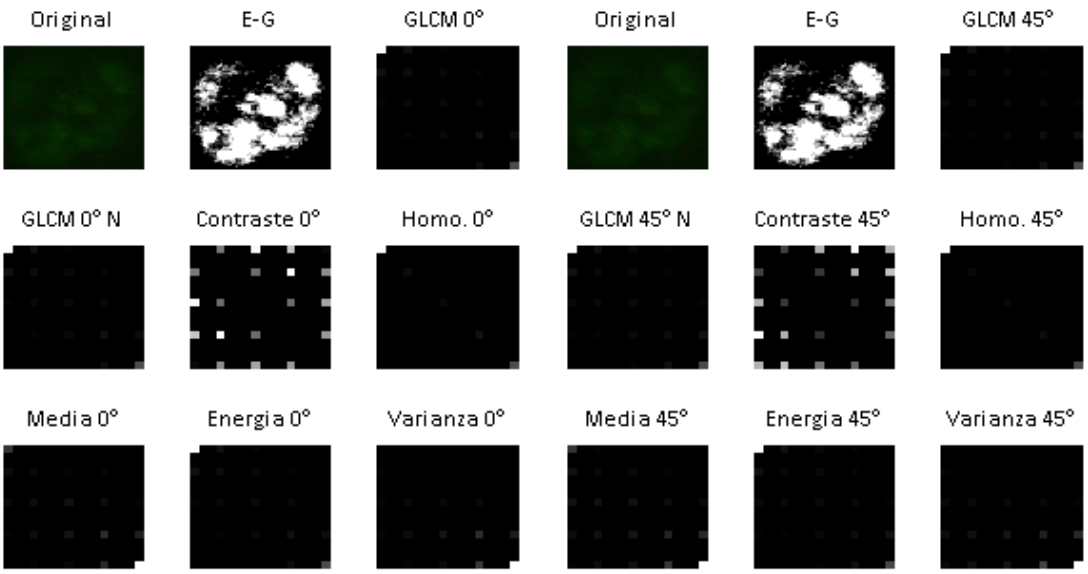


GLCM 90 °Moteada gruesa GLCM 135°Moteada gruesa

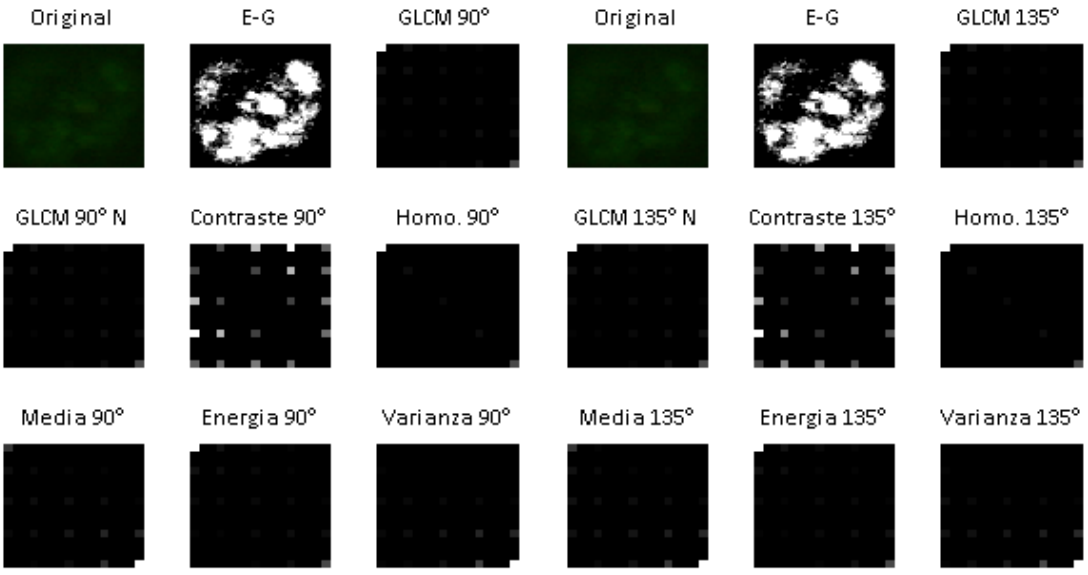
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.10: GLCM moteada gruesa (favorable) .

CÉLULA MOTEADA GRUESA (imagen desfavorable)



GLCM 0 °Moteada gruesa GLCM 45 °Moteada gruesa

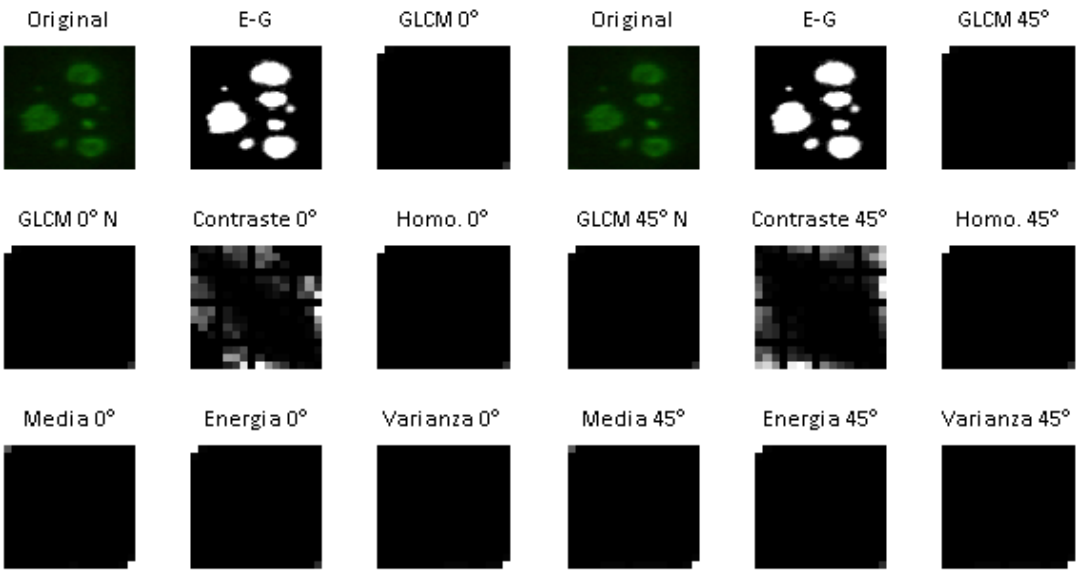


GLCM 90 °Moteada gruesa GLCM 135°Moteada gruesa

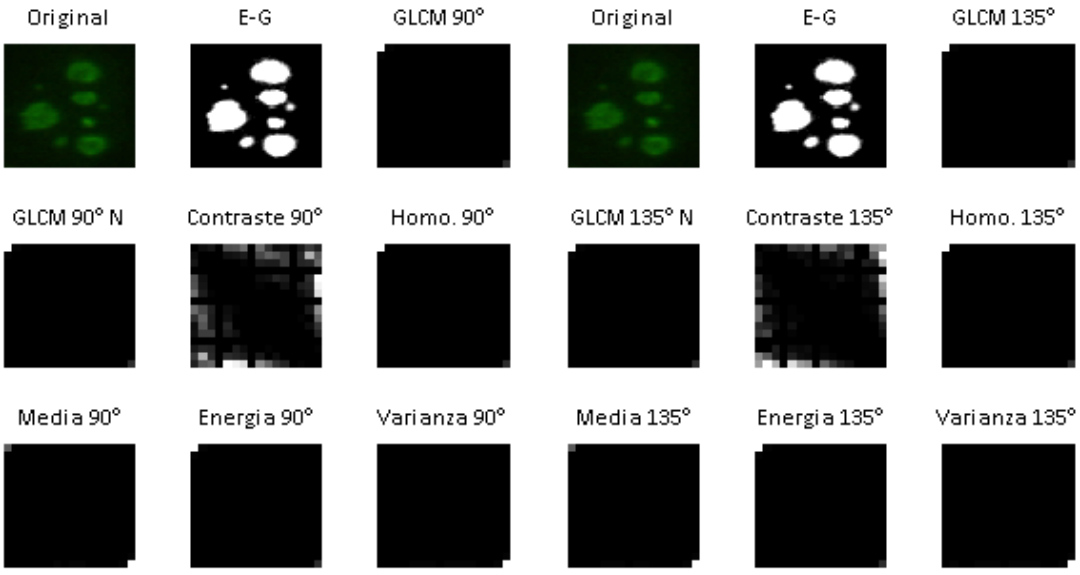
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.11: GLCM moteada gruesa (desfavorable) .

CÉLULA NUCLEOLAR (imagen favorable)



GLCM 0 °Nucleolar GLCM 45 °Nucleolar

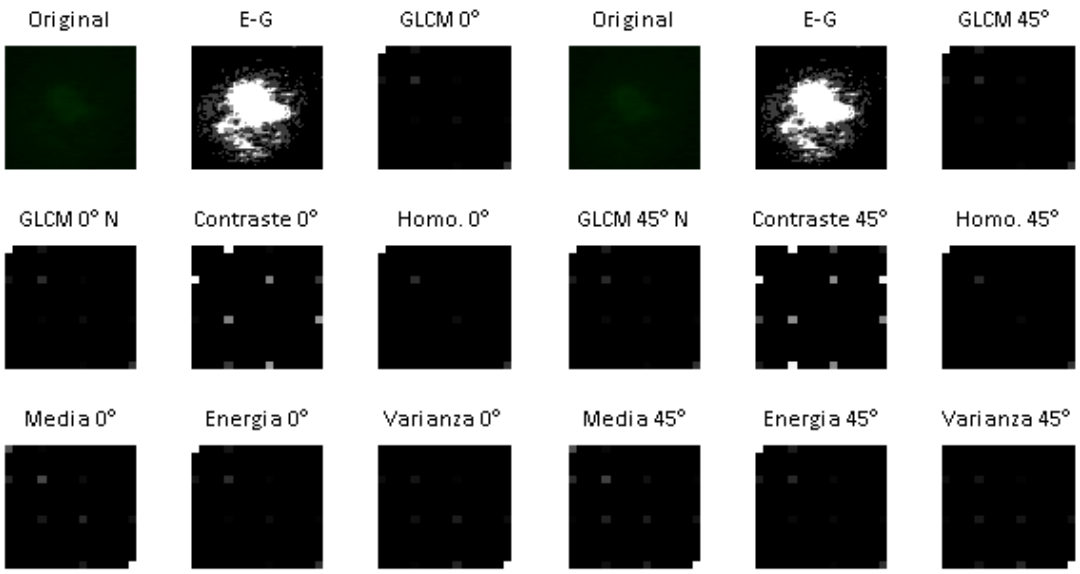


GLCM 90 °Nucleolar GLCM 135°Nucleolar

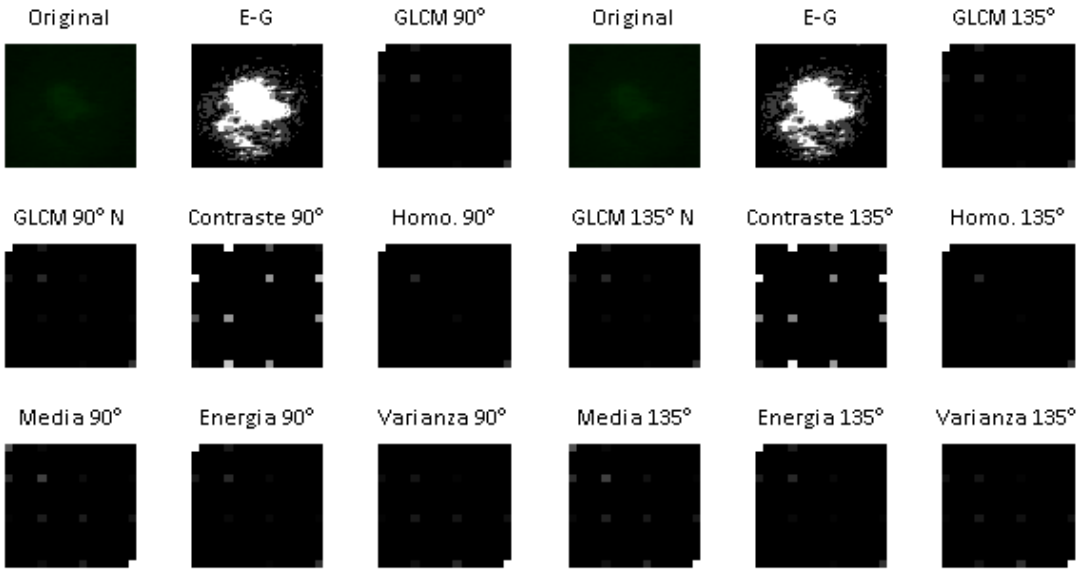
* E - G (Escala de Grises) y N (Normalizada).

Figura 4.12: GLCM nucleolar (favorable) .

CÉLULA NUCLEOLAR (imagen desfavorable)



GLCM 0 °Nucleolar GLCM 45 °Nucleolar



GLCM 90 °Nucleolar GLCM 135°Nucleolar

* E - G (Escala de Grises) y N (Normalizada).

Figura 4.13: GLCM nucleolar (desfavorable) .

4.2. Matriz de Descomposición Wavelet 2D

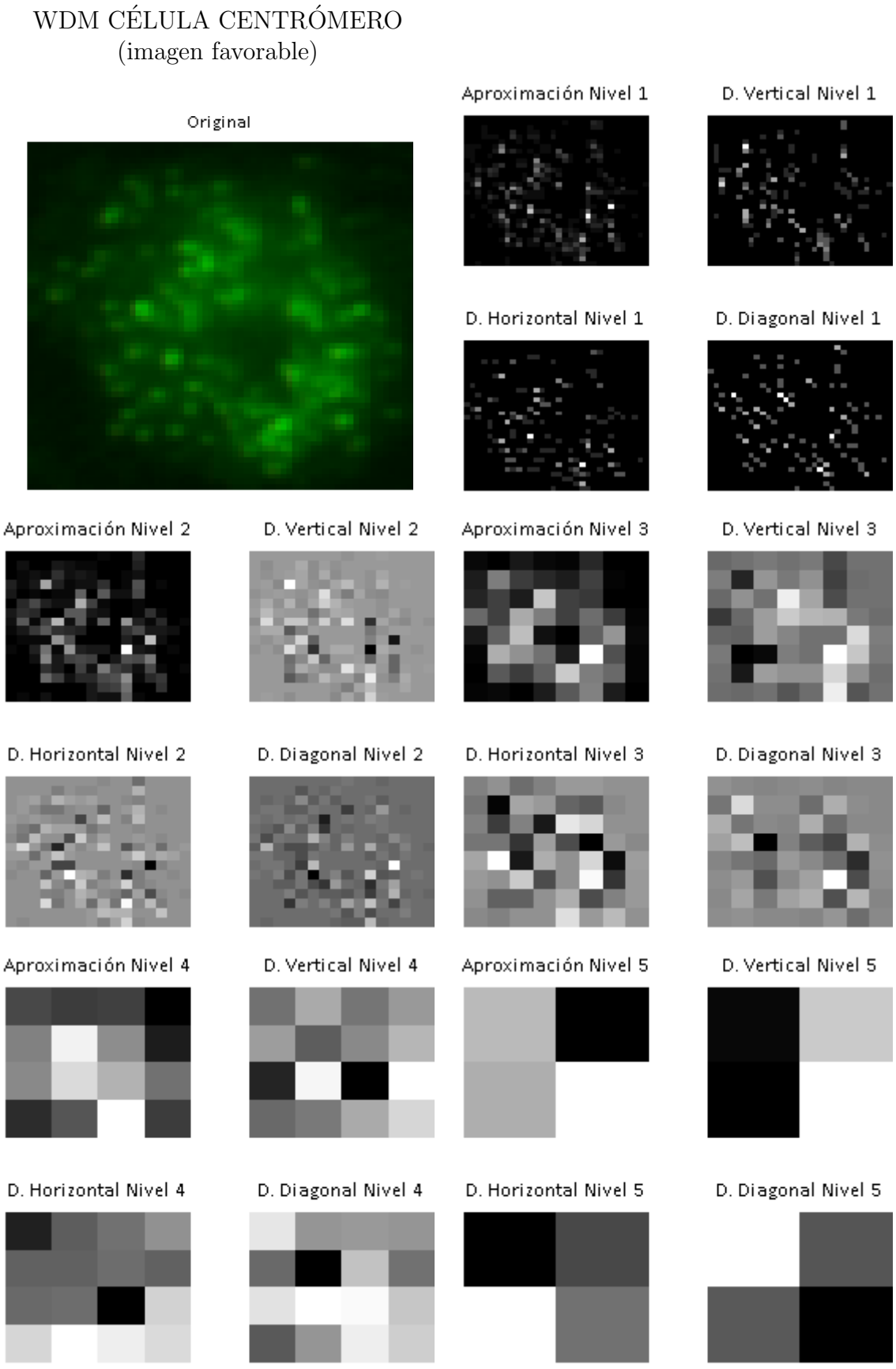


Figura 4.14: WDM centrómero (favorable) .

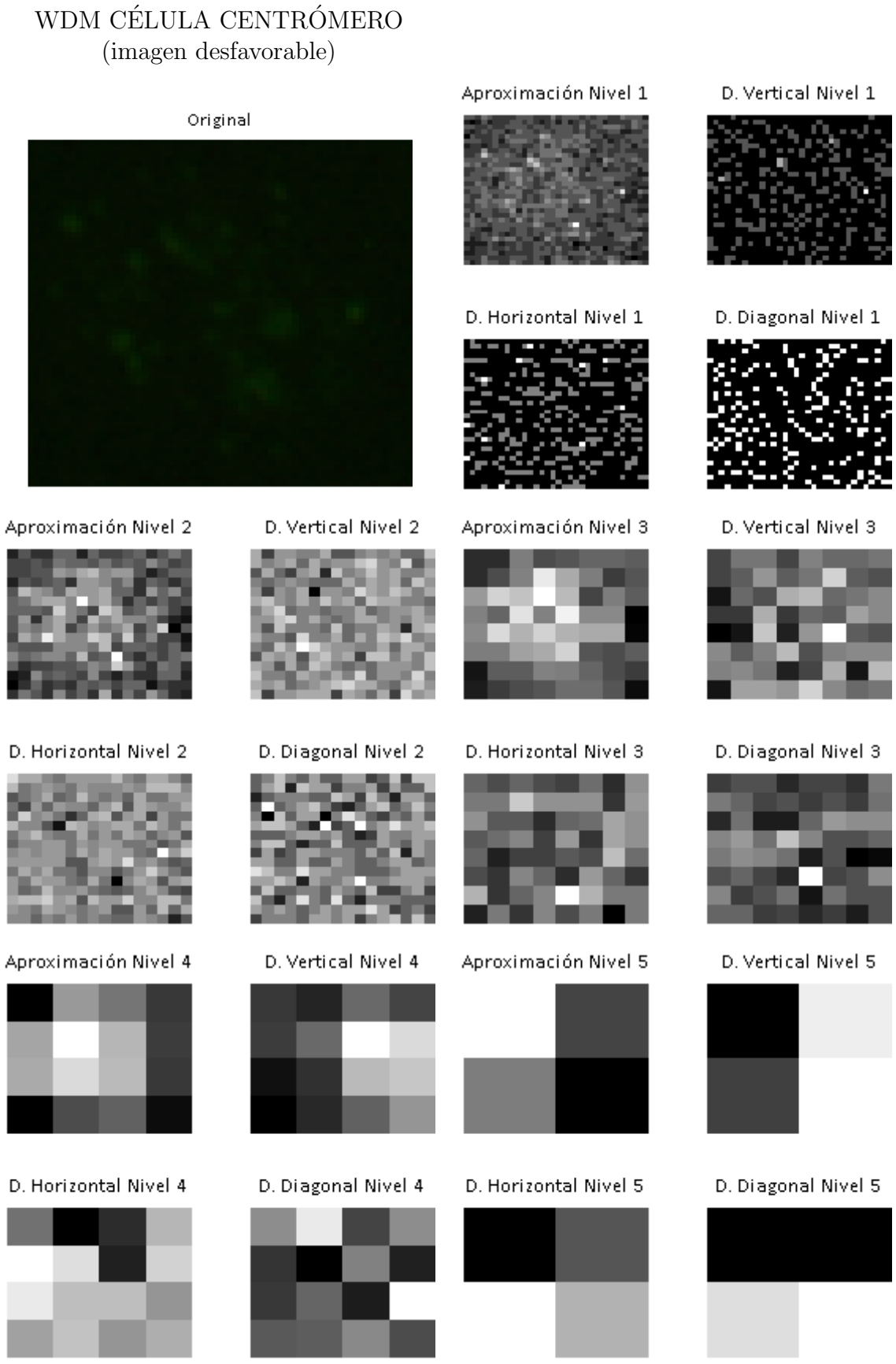


Figura 4.15: WDM centrómero (desfavorable) .

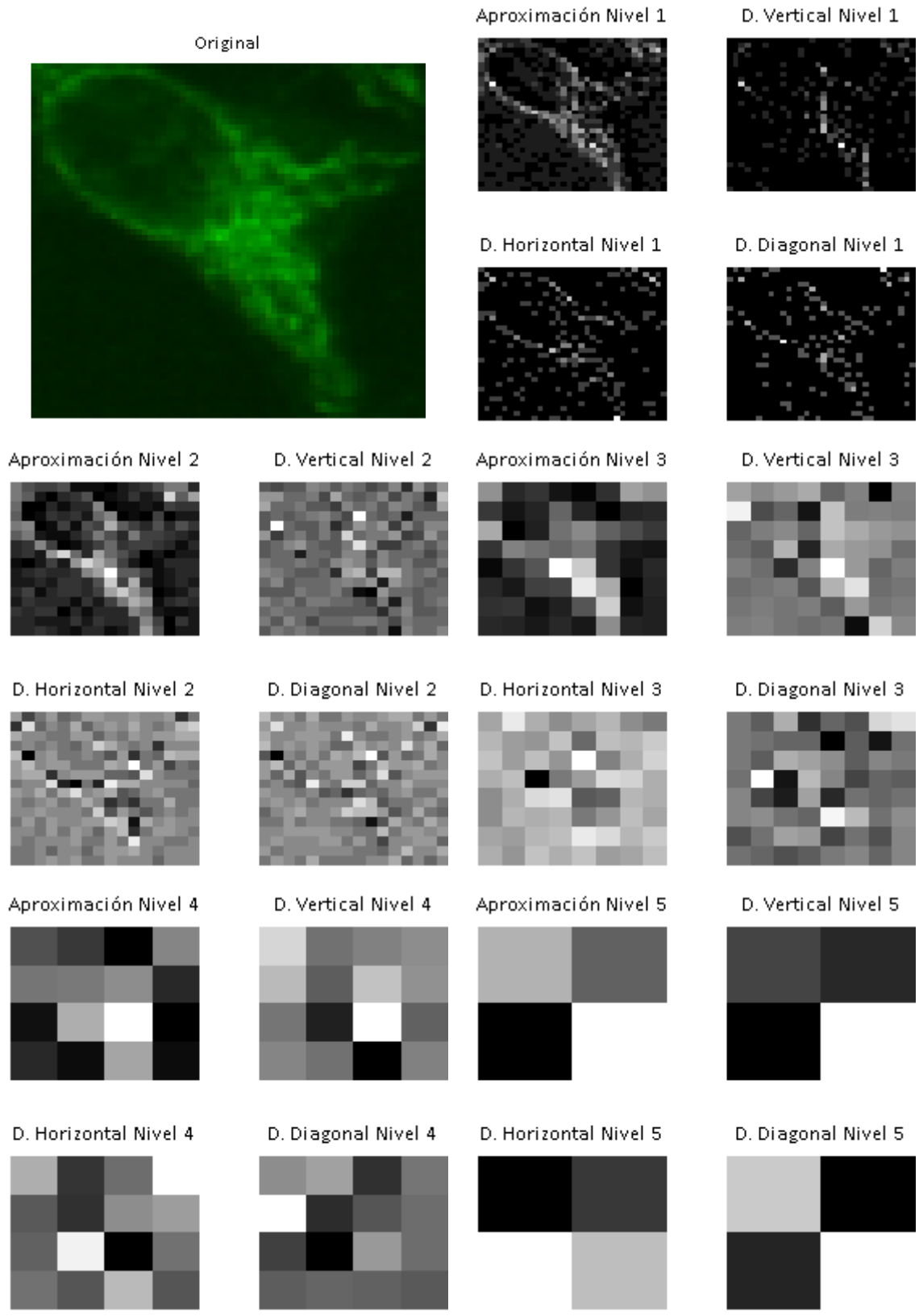


Figura 4.16: WDM citoplasmático (favorable) .

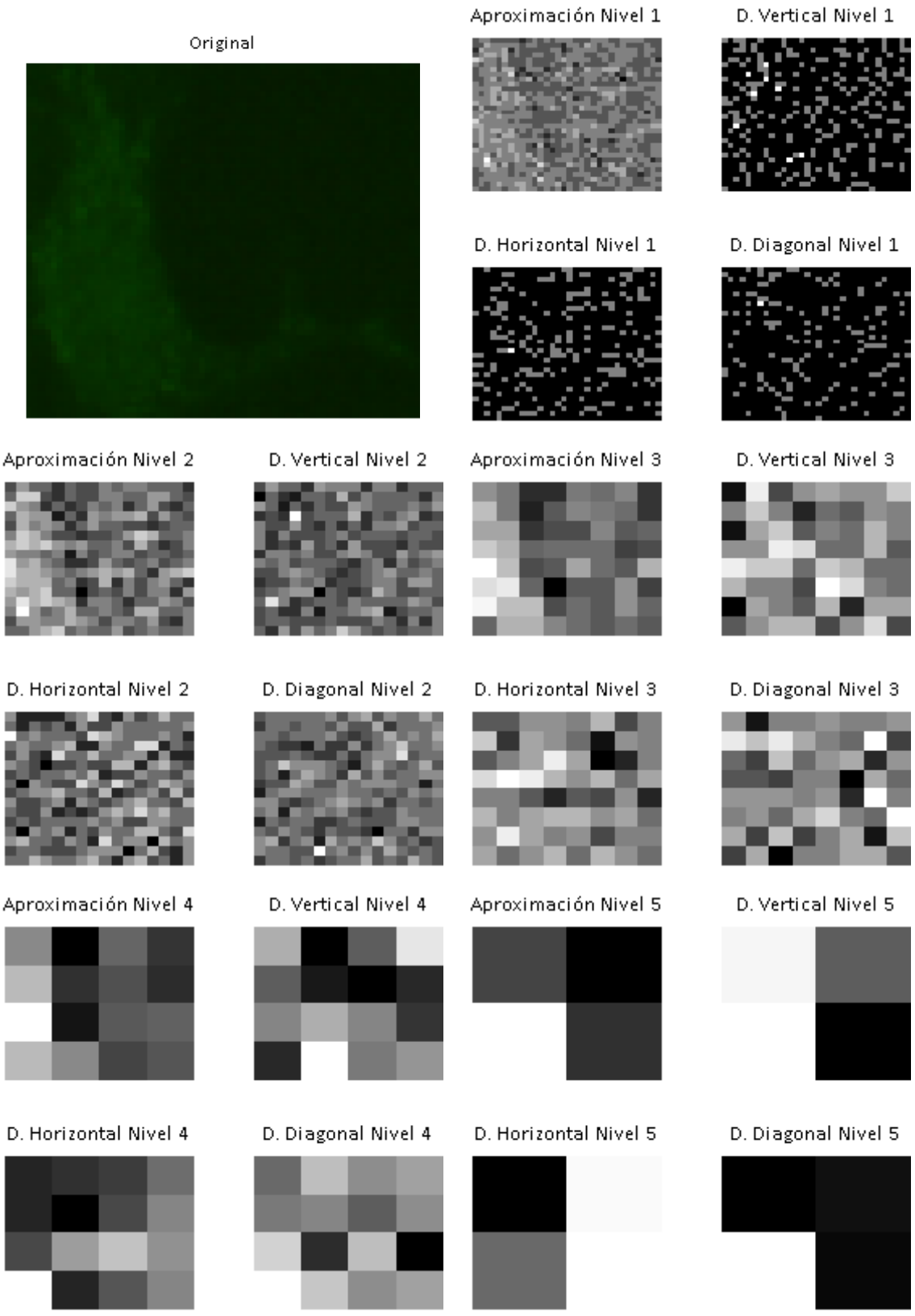


Figura 4.17: WDM citoplasmático (desfavorable) .

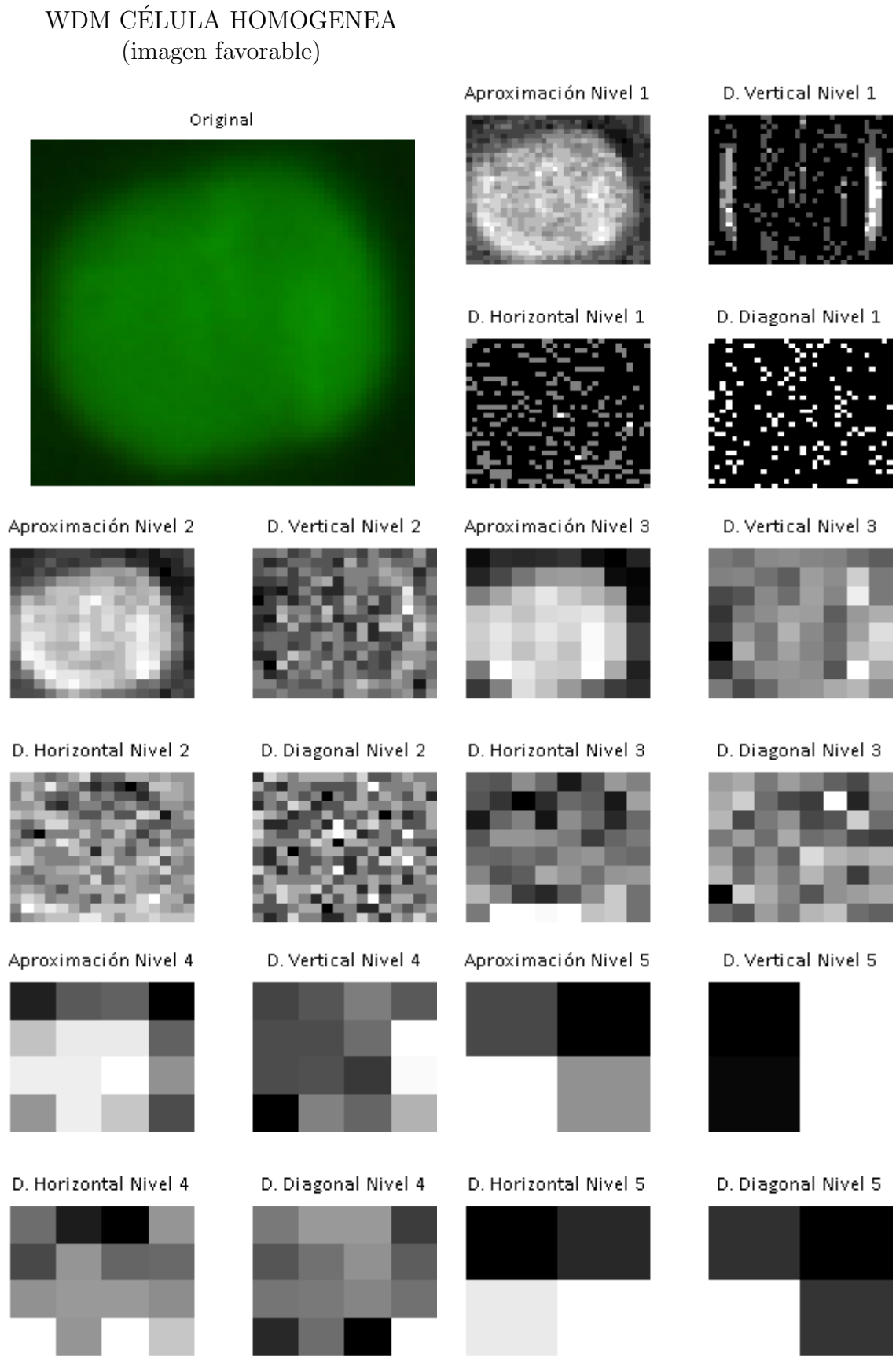


Figura 4.18: WDM homogénea (favorable) .

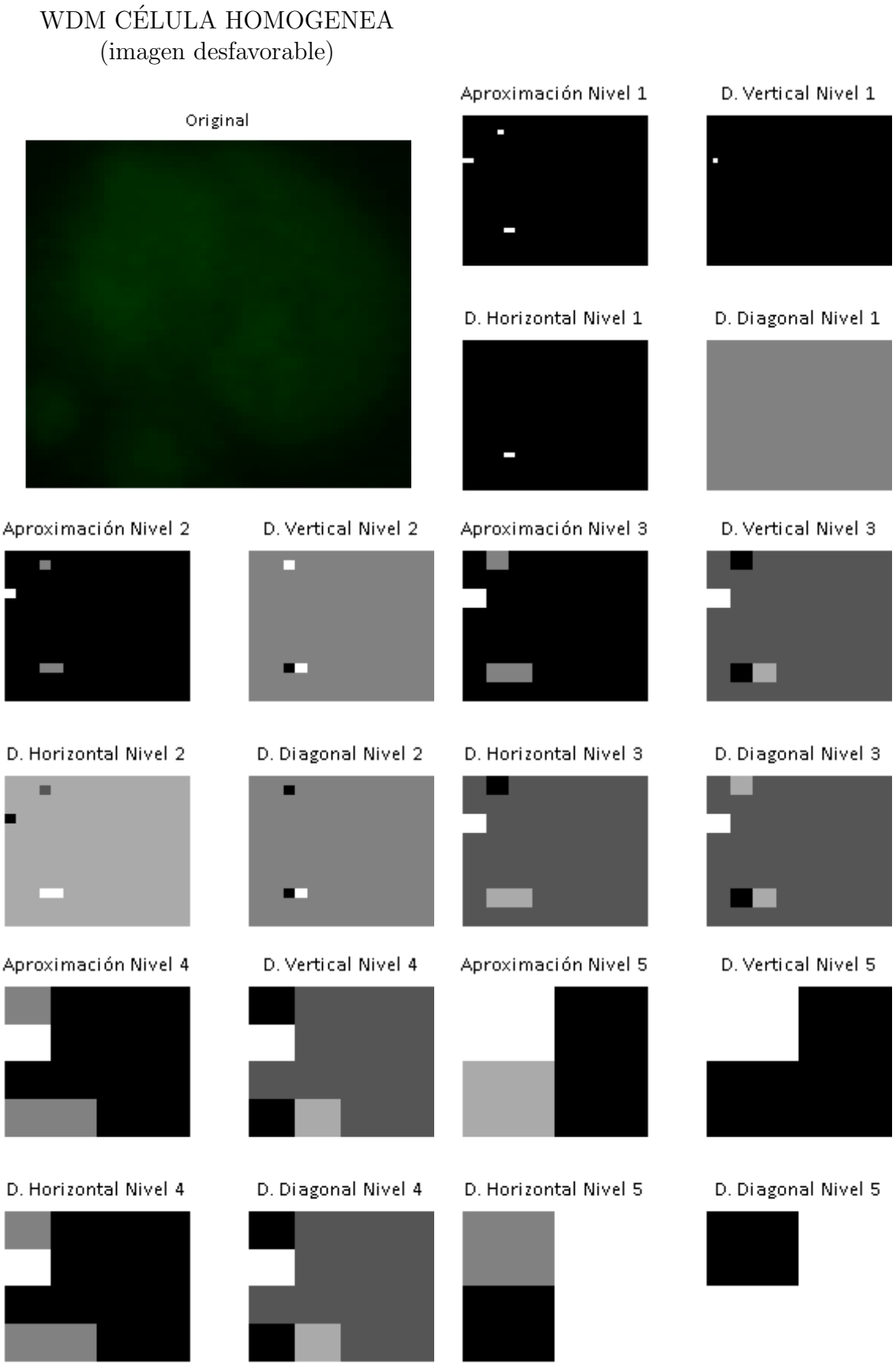


Figura 4.19: WDM homogénea (desfavorable) .

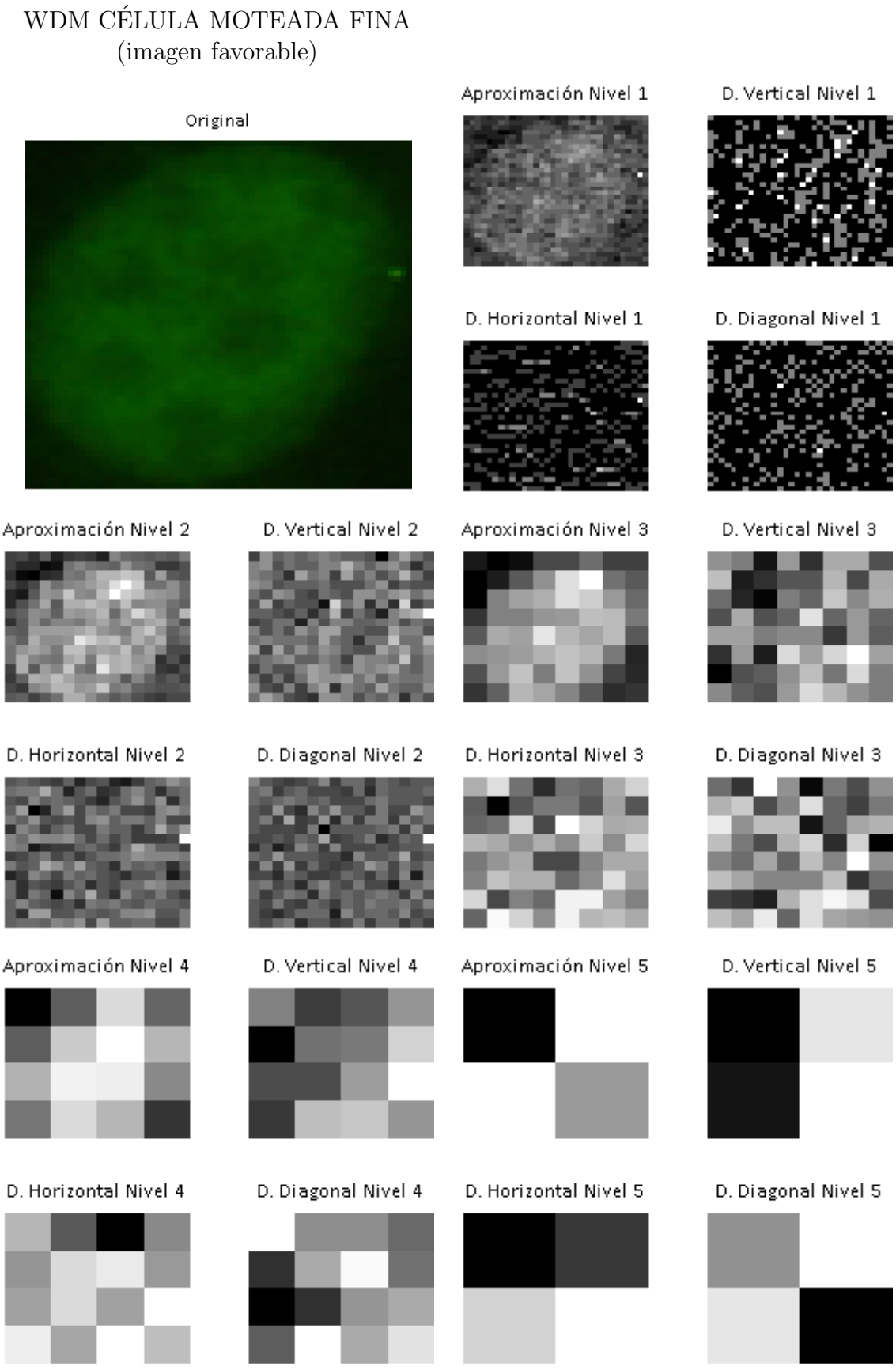


Figura 4.20: WDM moteada fina (favorable) .

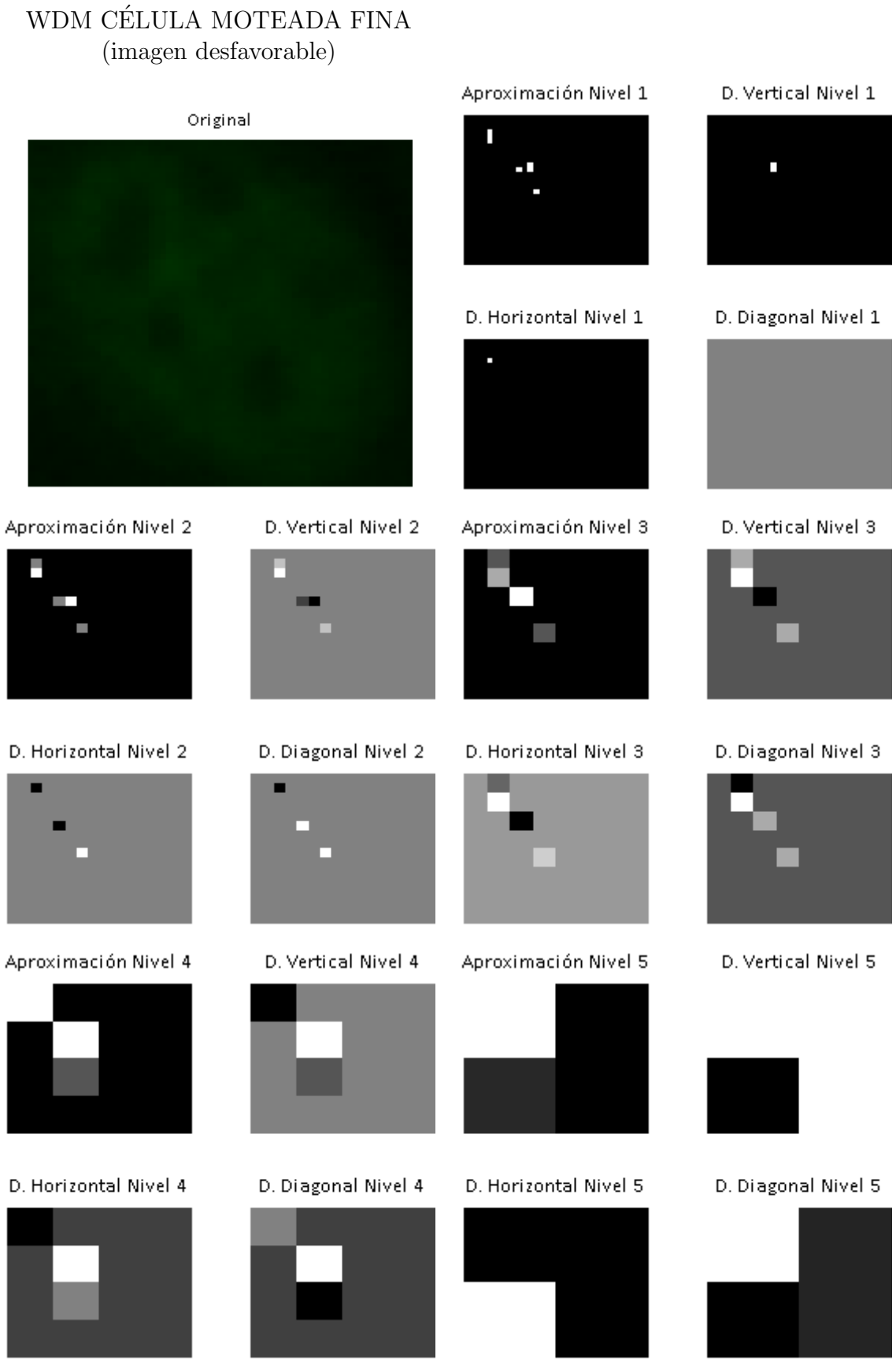


Figura 4.21: WDM moteada fina (desfavorable) .

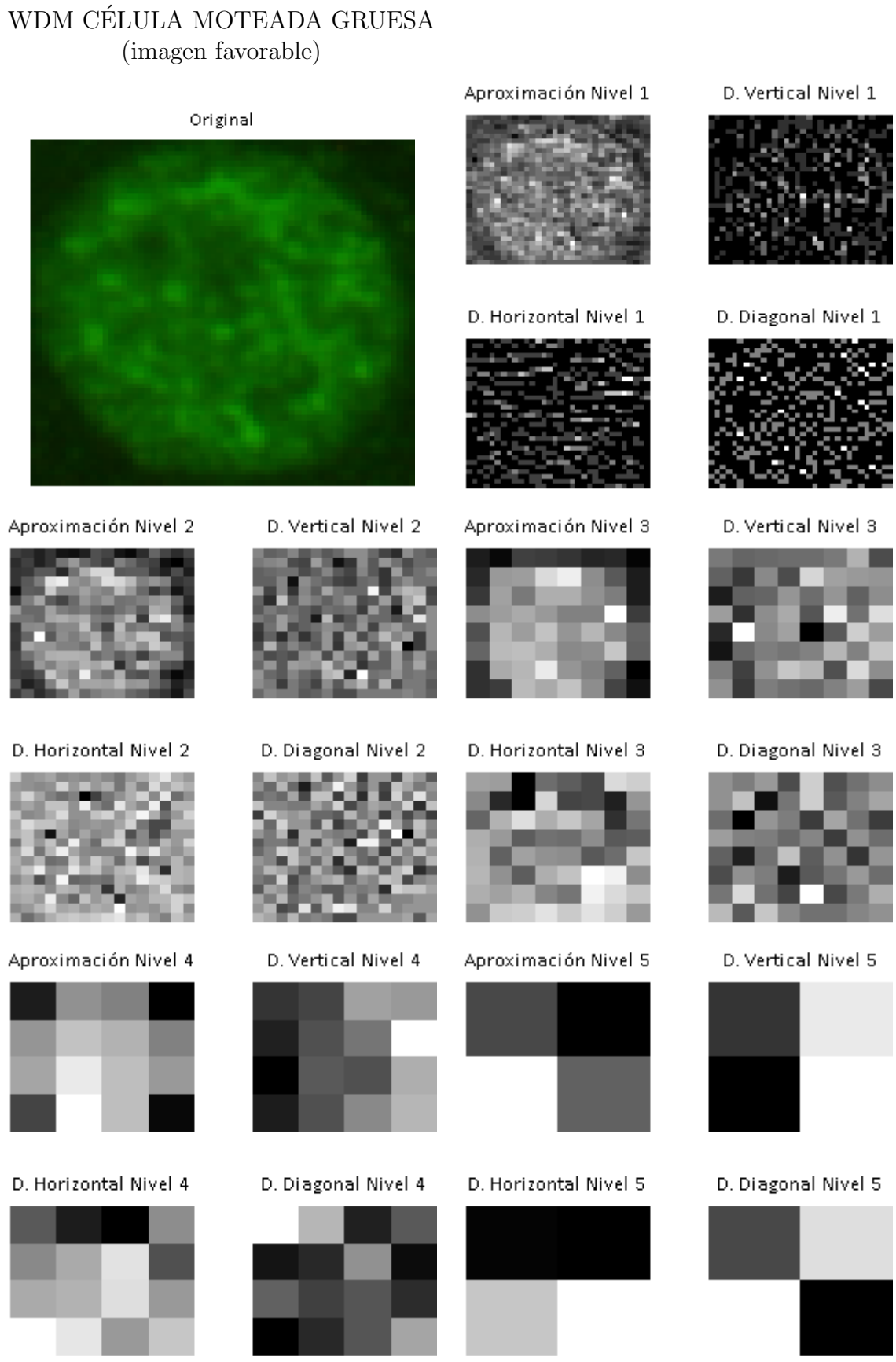


Figura 4.22: WDM moteada gruesa (favorable) .

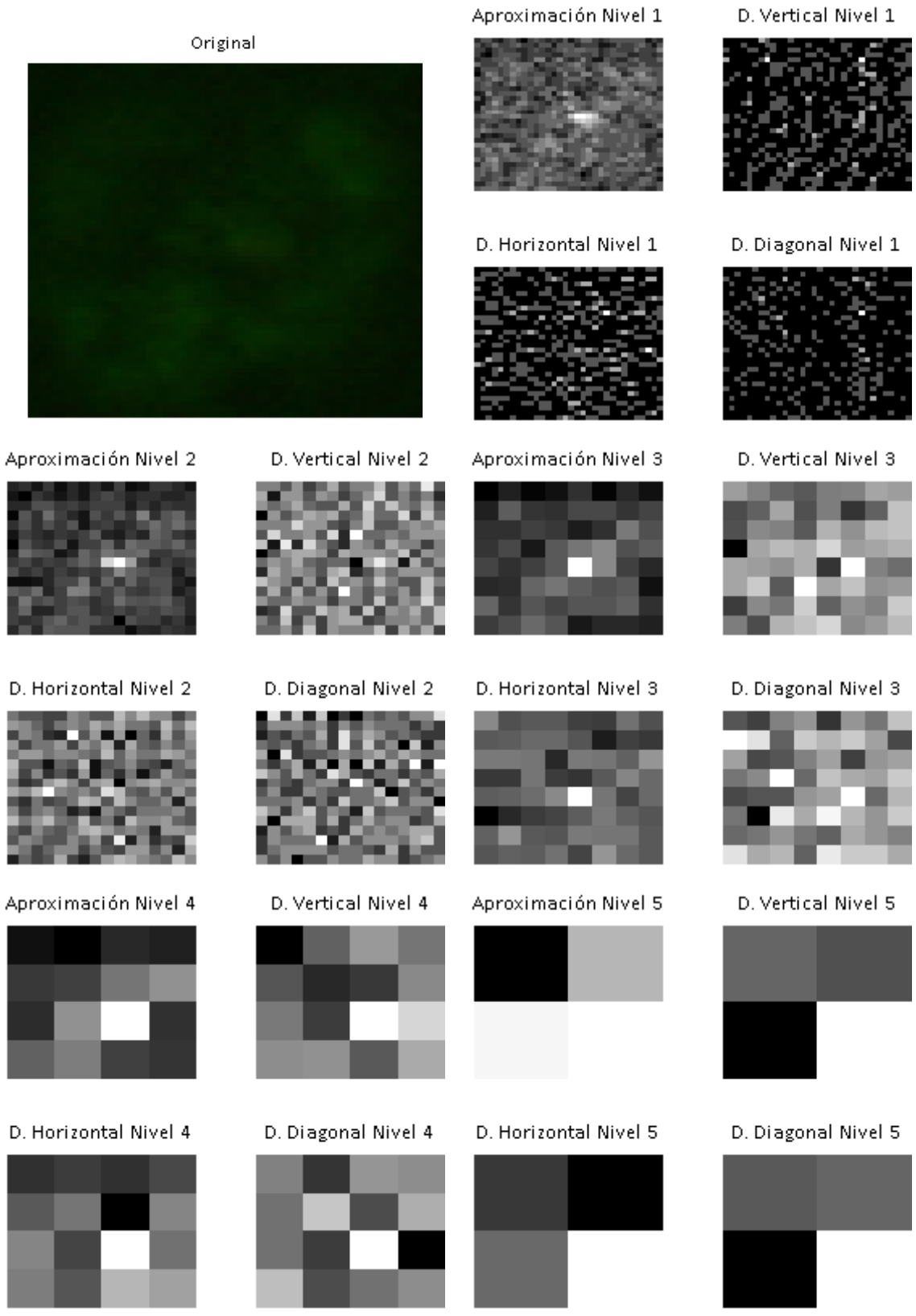


Figura 4.23: WDM moteada gruesa (desfavorable) .

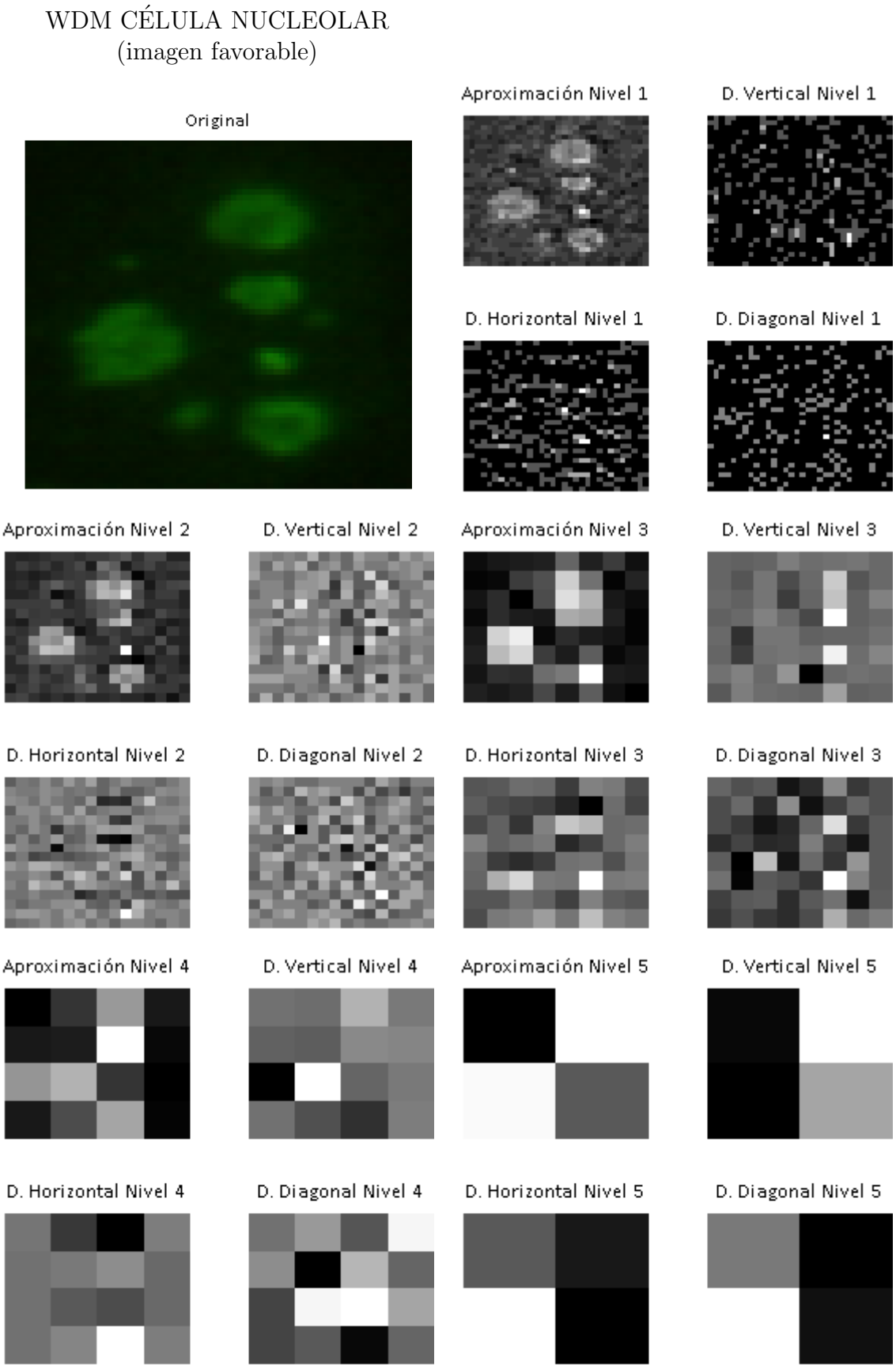


Figura 4.24: WDM nucleolar (favorable) .

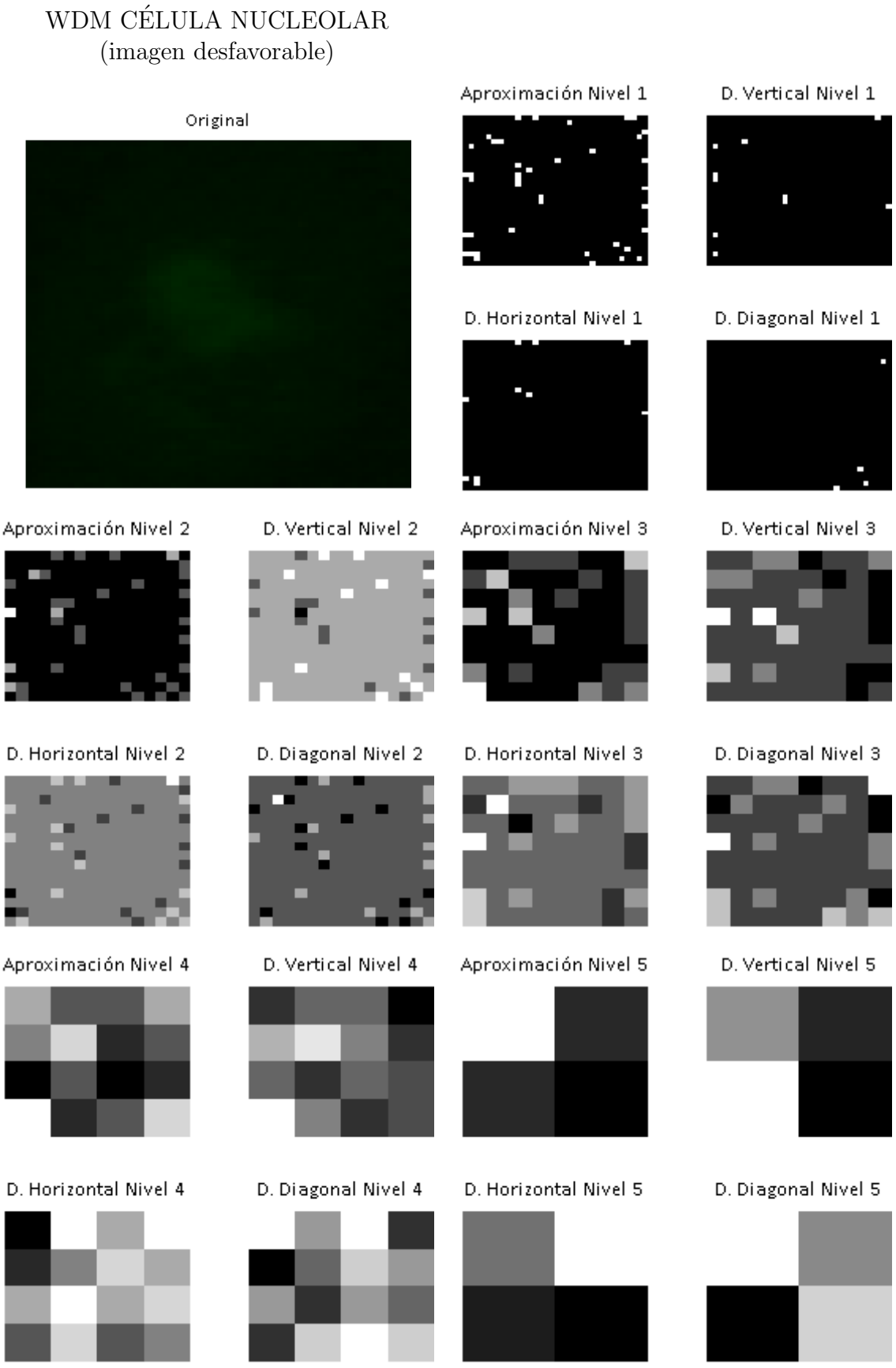


Figura 4.25: WDM nucleolar (desfavorable) .

4.3. Tablas de Resultados y Gráficas

A continuación se muestran los resultados obtenidos al aplicar ambos métodos de extracción de características sobre tres tipos de conjuntos de datos: un conjunto de 120 imágenes (20 por cada clase), 300 imágenes (50 imágenes por cada clase) y 721 imágenes (diferente número de imágenes por clase) y aplicando el Árbol de Clasificación y Regresión como clasificador.

En el caso de las GLCM primero se convirtió la imagen a escala de grises para después ser normalizada y extraer las características, se agregó como característica adicional la desviación estándar a la imagen y la media de la desviación estándar a cada direccional de la imagen, donde se comprueba que el mejor resultado se obtiene al utilizar las seis características base y la desviación estándar de la imagen.

Se hicieron pruebas en diferentes dimensiones de la imagen (64x64 px y 128x128 px) en donde el resultado más preciso se obtuvo con imágenes de 64x64 px con todo el conjunto de imágenes anteriormente descrito.

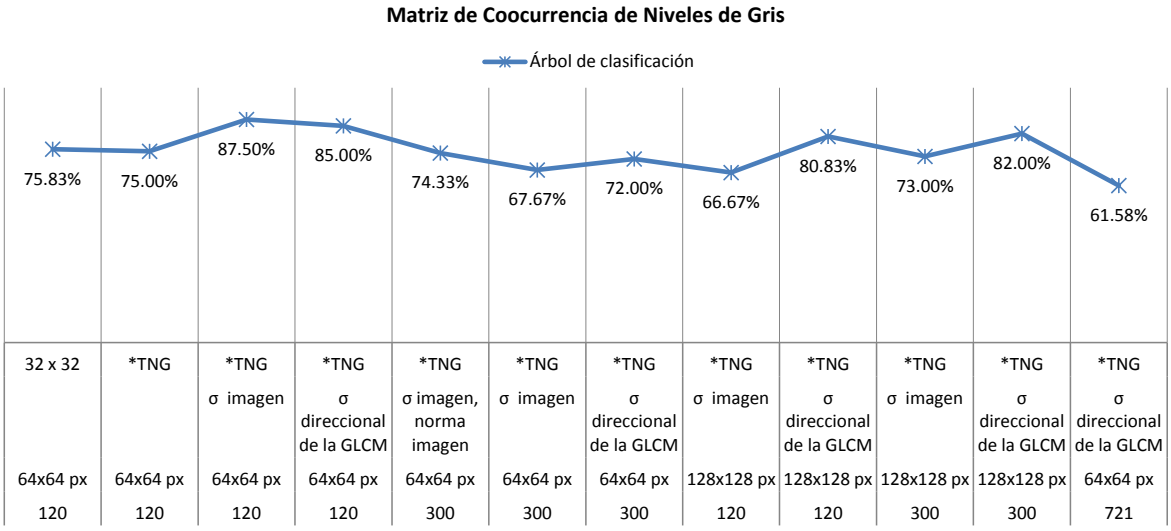
En el caso de la WDM de 2 dimensiones se utilizo imágenes de 64x64 px para lograr una homogeneidad en ambos casos y se extrajo características en todos los niveles, en primera instancia para clasificar se tomó en cuenta características por cada nivel, después las características por nivel agregándole la media y desviación estándar del componente de aproximación, y por último todas las características de todos los niveles, la media y desviación estándar del componente de aproximación, haciendo un total de 32 características.

Se observa que la mejor clasificación se obtiene al hacer uso de las 32 características, después con las ocho características por nivel y por último con las seis características por nivel, donde se aprecia que hay un decremento importante al usar estas seis características; por lo tanto las características de desviación estándar y la media del componente de aproximación son las que ayudan en gran medida a la clasificación con el Árbol de Clasificación y Regresión.

A continuación un reporte detallado sobre los resultados de clasficación con base en lo anteriormente descrito.

4.3.1. Gráficas y tablas GLCM

El cuadro 4.1 describe las distintas pruebas realizadas a distintos niveles de grises: 32 niveles y todos los niveles de gris (* TNG.), características propias de la GLCM (X) más las características adicionales (Y): desviación estándar de la imagen, desviación estándar de la imagen + norma de la imagen, y desviación estándar a cada direccional de la imagen(en esta última se calculó la desviación estándar sobre las 4 respectivas direccionales y posteriormente se promediaron), resoluciones de la imagen de distinto tamaño: 64x64 pixeles y 128x128 pixeles, y cantidades distintas del número de imágenes en las pruebas: 120, 300 y 721 (para el caso de 120 imágenes 20 le corresponde a cada clase, 300 imágenes le corresponde 50 imágenes por clase, y 721 donde se describe el número de clases ocupadas por clase en el cuadro 4.3). Se observa que la mejor clasificación en el Árbol de Clasificación y Regresión con base en las características descritas anteriormente se obtiene el mejor resultado al hacer uso de 120 imágenes, con todos los niveles de grises, sobre una imagen de 64x64 pixeles y agregando como característica la desviación estándar a la imagen obteniendo un 87.5 % de clasificados exitosamente. El cuadro 4.3 describe los datos ocupados para dicha gráfica.



Cuadro 4.1: Gráfica comparativa: distintos niveles de gris, características adicionales, diferentes resoluciones.

Característica	Descripción
Q	Clases de células : Q ∈ { CE, CI, HO, MG, MF, NU }
X	Características: X∈ { C,H,M,N,V}
Y	Características adicionales: Y ∈ { DEI, DEDI, NI}
CE	Clase célula centrómero
CI	Clase célula citoplasmático
HO	Clase célula homogénea
MG	Clase célula moteada gruesa
MF	Clase célula moteada fina
NU	Clase célula nucleolar
C	GLCM contraste
H	GLCM homogeneidad
M	GLCM media
N	GLCM energía
V	GLCM varianza
DEI	GLCM desviación estándar de la imagen
DEDI	GLCM desviación estándar direccional de la imagen
NI	GLCM norma de la imagen
W_n	Coeficiente wavelet combinado en nivel n

Cuadro 4.2: Reporte de características usadas en las pruebas.

El siguiente cuadro describe las distintas pruebas mencionadas en el cuadro 4.1, manteniendo las características propias de la GLCM X mas las adicionales Y, con base en las clases de células Q en la cual podemos observar el porcentaje, para cada clase, de imágenes clasificadas correctamente. Además, para el Árbol de Clasificación y Regresión las clases que obtienen el 100 % de imágenes clasificadas correctamente son la Moteada Gruesa, Moteada Fina, y Homogenea sobre un conjunto de prueba de 120 imágenes, a 64x64 pixeles agregando como característica adicional la desviación estándar de la imagen haciendo uso de todos los niveles de gris.

Se observa que la desviación estándar de la imagen tiene mayor influencia sobre el Árbol de Clasificación que la desviación estándar a las direccionales de la imagen sobre el mismo conjunto de condiciones, ya que existe una diferencia de 2.5 % entre ellas al realizar las pruebas.

					ÁRBOL DE CLASIFICACIÓN						
Total Img.	**Img. por Clase**	**Tam. Imagen**	**Características adicionales**	**Dim. GLCM**	**MG**	**MF**	**CE**	**HO**	**NU**	**CI**	**Total Clasificadas**
120	20	64x64 px		32 x 32	20	14	12	16	17	12	91
					100%	70%	60%	80%	85%	60%	75.83%
				*TNG	17	20	14	19	9	11	90
					85%	100%	70%	95%	45%	55%	75.00%
			σ imagen		20	20	14	20	13	18	105
					100%	100%	70%	100%	65%	90%	87.50%
			σ direccional de la GLCM		20	20	15	20	14	13	102
					100%	100%	75%	100%	70%	65%	85.00%
300	50		σ imagen, norma imagen		40	30	33	36	36	28	203
					80%	60%	66%	72%	72%	56%	67.67%
			σ imagen		37	38	33	40	38	30	216
					74%	76%	66%	80%	76%	60%	72.00%
			σ direccional de la GLCM		46	34	36	37	38	32	223
					92%	68%	72%	74%	76%	64%	74.33%
120	20	128x128 px	σ imagen		18	20	3	14	11	14	80
					90%	100%	15%	70%	55%	70%	66.67%
			σ direccional de la GLCM		18	20	11	15	15	18	97
					90%	100%	55%	75%	75%	90%	80.83%
300	50		σ imagen		50	37	33	44	31	24	219
					100%	74%	66%	88%	62%	48%	73.00%
			σ direccional de la GLCM		44	42	38	43	40	39	246
					88%	84%	76%	86%	80%	78%	82.00%
721	**	64x64 px	σ direccional de la GLCM		90	51	115	85	65	38	444
					83%	54%	55%	57%	64%	66%	61.58%

* TNG (Todos los Niveles de Gris)
** No.Img.por Clase

MG	MF	CE	HO	NU	CI	Total
109	94	208	150	102	58	721

Cuadro 4.3: GLCM. Tabla comparativa: distintos niveles de gris, características adicionales, diferentes resoluciones.

En el cuadro 4.4 se describe a detalle los porcentajes y números reales sobre las pruebas realizadas, donde se puede apreciar los máximos porcentajes con base en el número de imágenes y resolución.

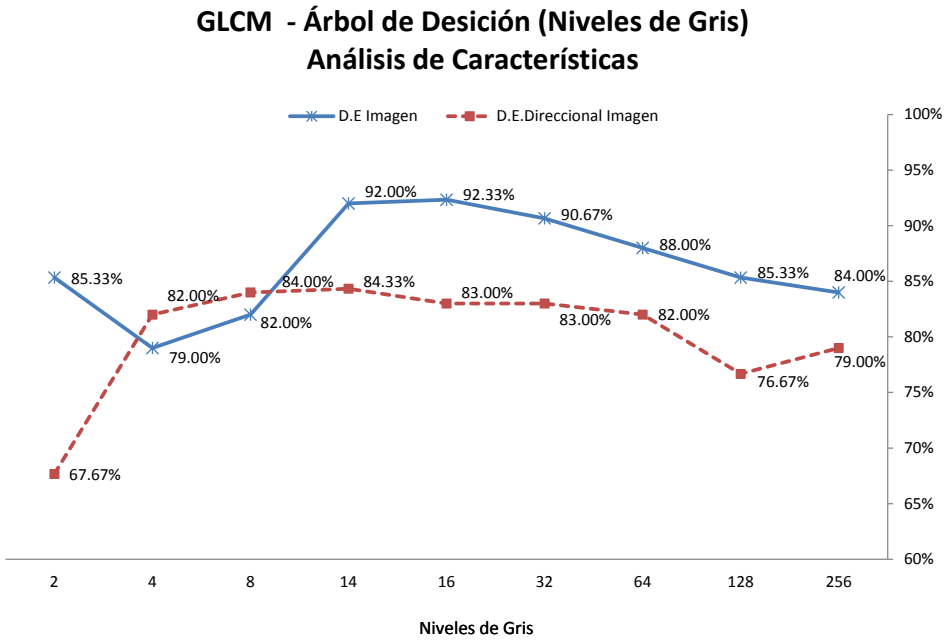
					ÁRBOL DE CLASIFICACIÓN						
Total Img.	**Img. por Clase**	**Tam. Imagen**	**Características adicionales**	**Dim. GLCM**	**MG**	**MF**	**CE**	**HO**	**NU**	**CI**	**Total Clasificadas**
120	20	64x64 px		32 x 32	20	14	12	16	17	12	91
					100%	70%	60%	80%	85%	60%	75.83%
				*TNG	17	20	14	19	9	11	90
					85%	100%	70%	95%	45%	55%	75.00%
			σ imagen		20	20	14	20	13	18	105
					100%	100%	70%	100%	65%	90%	87.50%
			σ direccional de la GLCM		20	20	15	20	14	13	102
					100%	100%	75%	100%	70%	65%	85.00%
300	50		σ imagen, norma imagen		40	30	33	36	36	28	203
					80%	60%	66%	72%	72%	56%	67.67%
			σ imagen		37	38	33	40	38	30	216
					74%	76%	66%	80%	76%	60%	72.00%
			σ direccional de la GLCM		46	34	36	37	38	32	223
					92%	68%	72%	74%	76%	64%	74.33%
120	20	128x128 px	σ imagen		18	20	3	14	11	14	80
					90%	100%	15%	70%	55%	70%	66.67%
			σ direccional de la GLCM		18	20	11	15	15	18	97
					90%	100%	55%	75%	75%	90%	80.83%
300	50		σ imagen		50	37	33	44	31	24	219
					100%	74%	66%	88%	62%	48%	73.00%
			σ direccional de la GLCM		44	42	38	43	40	39	246
					88%	84%	76%	86%	80%	78%	82.00%
721	**	64x64 px	σ direccional de la GLCM		90	51	115	85	65	38	444
					83%	54%	55%	57%	64%	66%	61.58%

* TNG (Todos los Niveles de Gris)

** No.Img.por Clase

MG	MF	CE	HO	NU	CI	Total
109	94	208	150	102	58	721

Cuadro 4.4: GLCM. Análisis a detalle del cuadro 4.3.



Cuadro 4.5: GLCM. Gráfica del Árbol de Clasificación y Regresión : Análisis de características a distintos niveles de gris

En el cuadro 4.5 se comparan a distintos niveles de gris las características adicionales que son la desviación estándar y la desviación estándar a la direccional de la imagen sobre un conjunto de 300 imágenes, con resolución de 64x64 px. En la cual se obtiene el más alto porcentaje de ambas características al usar 14 niveles de gris 92 % y 84.33 % respectivamente. El máximo porcentaje es obtenido con la desviación estándar de la imagen a 16 niveles de gris con un 92.33 %, con lo cual 16 de niveles es un tamaño adecuado, ya que mientras menor sea el tamaño de la matriz de co-ocurrencia (16 x 16) serán más rápidos los cálculos a comparación de una matriz de 256 x 256. Por lo consiguiente para las pruebas siguientes se mantiene este margen de 16 niveles.

Los datos utilizados se muestran en el cuadro 4.6.

El cuadro 4.6 muestra a detalle los números reales de las imágenes clasificadas exitosamente en el clasificador bajo un conjunto de 300 imágenes de 64x64 px y 128x128 px, bajo distintos niveles de gris y características ocupadas.

TONOS DE GRISES										
				ÁRBOL DE CLASIFICACIÓN						
Total Img	**Tam. Imagen**	**Características adicionales**	**Niveles gris**	**MG**	**MF**	**CE**	**HO**	**NU**	**CI**	**Total Clasificadas**
300	64x64 px	σ imagen	2	49	38	38	47	48	36	256
		σ direccional de la GLCM	2	24	30	35	38	45	31	203
		σ imagen	4	50	37	36	40	37	37	237
		σ direccional de la GLCM	4	48	40	37	39	45	37	246
		σ imagen	8	47	39	40	40	46	34	246
		σ direccional de la GLCM	8	46	44	41	44	44	33	252
		σ imagen	14	50	48	39	48	49	42	276
		σ direccional de la GLCM	14	47	42	39	49	46	30	253
		σ imagen	16	50	47	43	47	46	44	277
		σ direccional de la GLCM	16	50	44	37	44	46	28	249
		σ imagen	32	50	44	41	45	49	43	272
		σ direccional de la GLCM	32	50	49	33	43	46	28	249
		σ imagen	64	50	50	34	48	43	39	264
		σ direccional de la GLCM	64	50	48	35	45	42	26	246
		σ imagen	128	50	44	34	47	43	38	256
		σ direccional de la GLCM	128	49	44	30	35	43	29	230
		σ imagen	256	49	47	31	43	43	39	252
		σ direccional de la GLCM	256	49	48	35	34	43	28	237
	128x128 px	σ imagen	16	45	41	40	46	45	40	257
		σ direccional de la GLCM	16	47	44	35	41	45	39	251
		σ imagen	32	46	45	39	45	47	31	253
		σ direccional de la GLCM	32	46	45	34	43	45	35	248

Cuadro 4.6: GLCM. Tabla de imágenes clasificadas en el Árbol de Clasificación y Regresión.

El cuadro 4.7 muestra en porcentajes los datos del cuadro 4.6 de ambos clasificadores para una mejor comprensión a nivel porcentual.

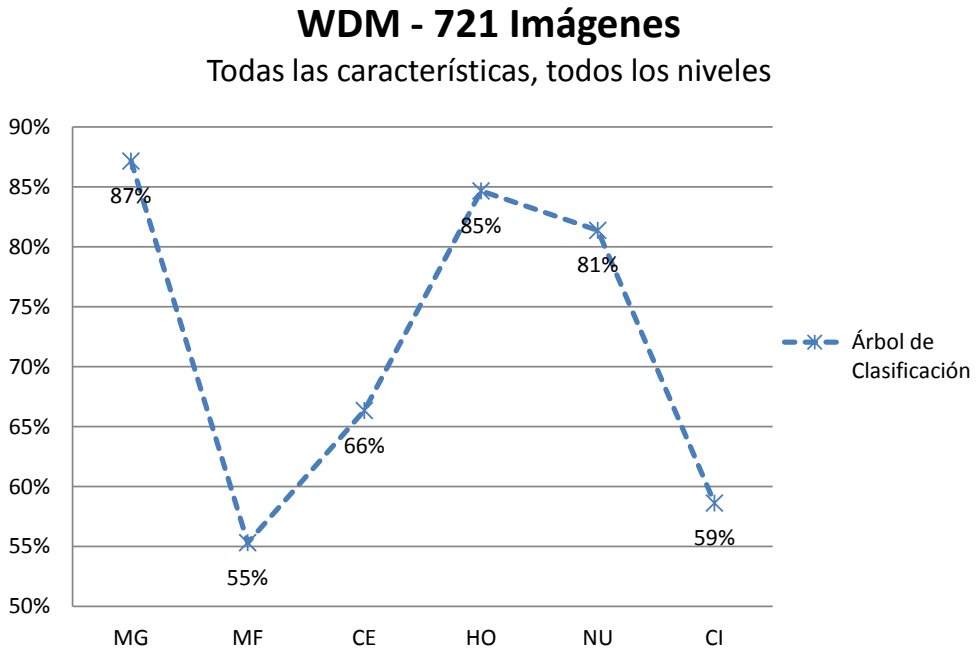
TONOS DE GRISES										
				ÁRBOL DE CLASIFICACIÓN						
Total Img	**Tam. Imagen**	**Características adicionales**	**Niveles gris**	**MG**	**MF**	**CE**	**HO**	**NU**	**CI**	**Total Clasificadas**
300	64x64 px	σ imagen	2	98%	76%	76%	94%	96%	72%	85.33%
		σ direccional de la GLCM	2	48%	60%	70%	76%	90%	62%	67.67%
		σ imagen	4	100%	74%	72%	80%	74%	74%	79.00%
		σ direccional de la GLCM	4	96%	80%	74%	78%	90%	74%	82.00%
		σ imagen	8	94%	78%	80%	80%	92%	68%	82.00%
		σ direccional de la GLCM	8	92%	88%	82%	88%	88%	66%	84.00%
		σ imagen	14	100%	96%	78%	96%	98%	84%	92.00%
		σ direccional de la GLCM	14	94%	84%	78%	98%	92%	60%	84.33%
		σ imagen	16	100%	94%	86%	94%	92%	88%	92.33%
		σ direccional de la GLCM	16	100%	88%	74%	88%	92%	56%	83.00%
		σ imagen	32	100%	88%	82%	90%	98%	86%	90.67%
		σ direccional de la GLCM	32	100%	98%	66%	86%	92%	56%	83.00%
		σ imagen	64	100%	100%	68%	96%	86%	78%	88.00%
		σ direccional de la GLCM	64	100%	96%	70%	90%	84%	52%	82.00%
		σ imagen	128	100%	88%	68%	94%	86%	76%	85.33%
		σ direccional de la GLCM	128	98%	88%	60%	70%	86%	58%	76.67%
		σ imagen	256	98%	94%	62%	86%	86%	78%	84.00%
		σ direccional de la GLCM	256	98%	96%	70%	68%	86%	56%	79.00%
	128x128 px	σ imagen	16	90%	82%	80%	92%	90%	80%	85.67%
		σ direccional de la GLCM	16	94%	88%	70%	82%	90%	78%	83.67%
		σ imagen	32	92%	90%	78%	90%	94%	62%	84.33%
		σ direccional de la GLCM	32	92%	90%	68%	86%	90%	70%	82.67%

Cuadro 4.7: GLCM. Tabla de los datos pertenecientes al cuadro 4.6 en porcentajes

4.3.2. Gráficas y tablas WDM

El cuadro 4.8 muestra el análisis del Árbol de Clasificación bajo un conjunto de 721 imágenes (MG=109, MF=94, CE=208, HO=150, NU=102, CI=58) de 64x64 pixeles.

Tomando en cuenta todas las características del vector W que lo conforman, 32 características contemplando todos los niveles: media del componente de aproximación, desviación estándar del componente de aproximación, y seis características propias de cada nivel, en este caso al ser imágenes de 64x64 pixeles se obtienen $p = 5$ niveles dando un total de $2 + 6p$ como se definió en la sección de la Transformada Wavelet.



Cuadro 4.8: WDM. Gráfica de imágenes clasificadas en el Árbol de Clasificación y Regresión (721 imágenes, todas las características, todos los niveles)

Como se observa bajo este conjunto de imágenes la clase mejor clasificada es la Moteada Gruesa con un 87 % seguida de la Homogénea con 85 %, secundada por la Nucleolar 81 %.

Los datos utilizados se muestran a continuación en el cuadro 4.9

Como se observa en el cuadro 4.9 se manejan imágenes de 64x64 pixeles y se toman a consideración el vector completo W de características. A pesar de ser un conjunto basto de 721 imágenes, y número de imágenes por clase desequilibrado se obtiene un 73 % de clasificados correctamente.

ARBOL CLASIFICACIÓN						
MG	MF	CE	HO	NU	CI	Total Imágenes
95	52	138	127	83	34	529
87%	55%	66%	85%	81%	59%	73.37%

Conjunto de 721 imágenes 64x64px, todas las características , todos los niveles

Cuadro 4.9: WDM. Tabla de imágenes clasificadas en el Árbol de Clasificación y Regresión (721 imágenes)

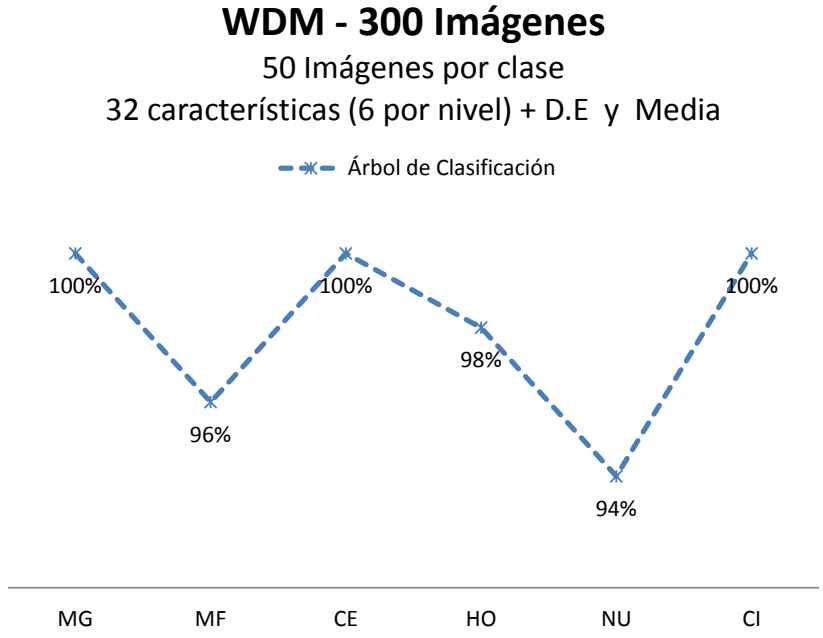
Se aprecia en el cuadro 4.10 que la colección de imágenes es de 300 y por cada clase le corresponde 50 imágenes con una dimensión de 64x64 pixeles, haciendo uso de las 32 características totales del vector W, 98 % de las imágenes son clasificadas correctamente en el Árbol de Clasificación y Regresión.

	ÁRBOL DE CLASIFICACIÓN						
Nivel	MG	MF	CE	HO	NU	CI	Total Imágenes
1-5	50	48	50	49	47	50	294
	100%	96%	100%	98%	94%	100%	98.00%

Conjunto de 50 imágenes por clase 64x64 px a color,
32 características (6 por nivel + Desviación Estándar y Media del componente aproximación).

Cuadro 4.10: WDM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes)

El cuadro 4.11 muestra el análisis del Árbol de Decisión al hacer uso para la prueba un conjunto de 50 imágenes por clase, teniendo un total de 300 imágenes de 64x64 pixeles, utilizando todas las características de todos los niveles (32 características) donde se puede observar que las imágenes exitosamente clasificadas pertenecen a la clase Moteada Gruesa, Centrómero y Citoplasmático con un 100 %.

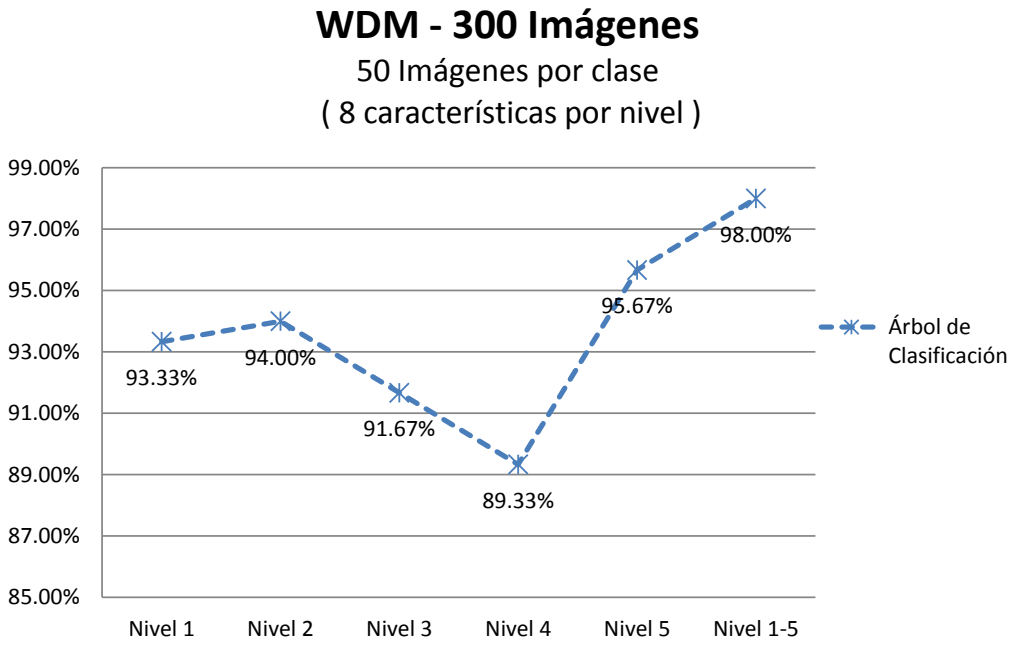


Cuadro 4.11: WDM. Gráfica de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes)

Los datos relacionados al cuadro 4.11 se encuentran en el cuadro 4.10 anteriormente descrito.

El cuadro 4.12 muestra el análisis del Árbol de Decisión a distintos niveles de descomposición (5 niveles) y adicionalmente los cinco niveles en conjunto, bajo un conjunto de 300 imágenes (50 por clase) de 64x64 pixeles y en cada nivel 8 características (desviación estándar del componente de aproximación + media del componente de aproximación + 6 características propias del nivel).

A partir del nivel 2 empieza a decaer el número de clasificados correctamente hasta llegar al 89.33 % en el nivel 4, ocurre un aumento radical en el nivel 5 donde supera la clasificación en los demás niveles con un 95.67 %. En el acumulado del nivel 1-5 (vector W completo) se logra el máximo porcentaje de clasificación de un 98 % convirtiéndolo en el más favorecedor.



Cuadro 4.12: WDM. Gráfica de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, 5 niveles, 8 características por nivel)

Los datos ocupados se encuentran en los cuadros 4.13 (Nivel: 1, 2, 3, 4, 5) y 4.14 (Nivel 1-5).

El cuadro 4.13 describe en número y porcentaje los datos clasificados del nivel 1 al nivel 5 para el clasificador sobre 50 imágenes por clase y dimensión de 64x64 pixeles. El máximo porcentaje se ubica en el nivel 5 con un 95.67 %, usando 8 características anteriormente descritas.

	ÁRBOL DE CLASIFICACIÓN						
Nivel	MG	MF	CE	HO	NU	CI	Total Imágenes
1	50	45	50	40	46	49	280
	100%	90%	100%	80%	92%	98%	93.33%
2	50	48	50	42	46	46	282
	100%	96%	100%	84%	92%	92%	94.00%
3	50	44	50	41	44	46	275
	100%	88%	100%	82%	88%	92%	91.67%
4	49	35	50	40	44	50	268
	98%	70%	100%	80%	88%	100%	89.33%
5	47	44	50	48	48	50	287
	94%	88%	100%	96%	96%	100%	95.67%

Conjunto de 50 imágenes por clase 64x64 px a color, 8 características en cada uno de los niveles.

Cuadro 4.13: WDM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, 5 niveles, 8 características por nivel)

El cuadro 4.14 muestra en número y porcentaje los datos clasificados usando las 32 características (correspondiente a todas las características del vector W). Se aprecia en el Árbol de Clasificación y Regresión que las clases que mejor clasifica son MG, CE y CI con un 100 %.

	ÁRBOL DE CLASIFICACIÓN						
Nivel	MG	MF	CE	HO	NU	CI	Total Imágenes
1-5	50	48	50	49	47	50	294
	100%	96%	100%	98%	94%	100%	98.00%

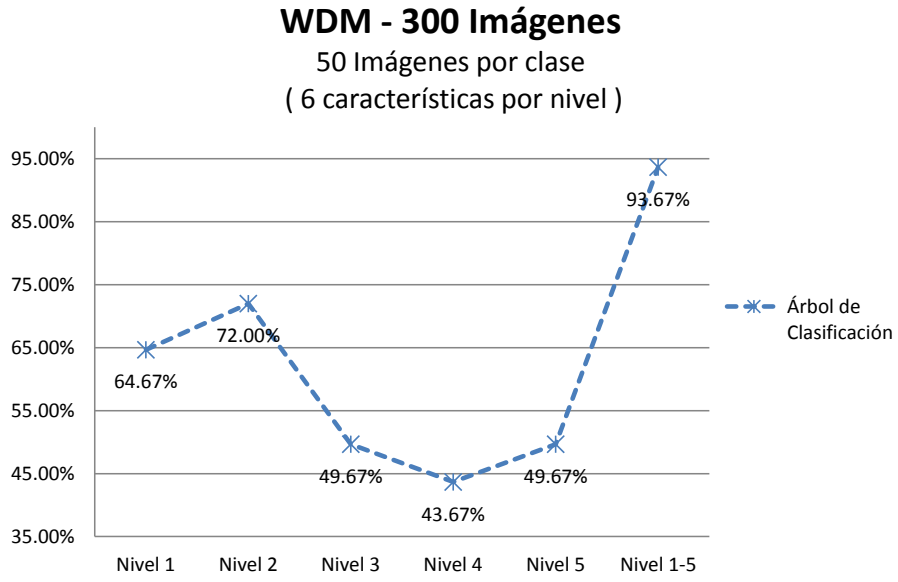
Conjunto de 50 imágenes por clase 64x64 px a color,
32 características (6 por nivel + Desviación Estándar y Media del componente aproximación).

Cuadro 4.14: WDM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, 32 características).

Cabe señalar que el conjunto de 32 características actua mejor sobre el conjunto del nivel 5 con 8 características, con una ventaja de 2.13 %.

El cuadro 4.15 muestra el análisis del Árbol de Clasificación a distintos niveles de descomposición (Nivel: 1, 2, 3, 4, 5 y Nivel Acumulado 1-5), bajo un conjunto de 300 imágenes (50 por clase) de 64x64 pixeles y en cada nivel 6 características propias de cada nivel (excluyendo la Desviación Estándar y la Media, ambas del componente de aproximación).

El éxito de clasificación se deteriora clasificando por independiente cada nivel, pero en conjunto todos los niveles alcanzan un 93.67 %.



Cuadro 4.15: WDM. Gráfica de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, 6 características por nivel)

Los datos se registran en los cuadros 4.16 (Nivel: 1, 2, 3, 4 y 5) y 4.17 (Nivel 1-5).

El cuadro 4.16 describe en número y porcentaje las imágenes clasificadas utilizando solo seis características. El Árbol de Clasificación y Regresión inicialmente clasifica con el 64 %, en el segundo nivel con el 72 % y a partir del tercero en adelante se decrementa el porcentaje, a lo cual las seis características no son suficientes para este método convirtiendo a las características que se excluyeron (media y desviación estándar, ambas del componente de aproximación) en fundamentales para la clasificación.

	ÁRBOL DE CLASIFICACIÓN						
Nivel	MG	MF	CE	HO	NU	CI	Total Imágenes
1	42	34	33	25	30	30	194
	84%	68%	66%	50%	60%	60%	64.67%
2	43	37	26	36	32	42	216
	86%	74%	52%	72%	64%	84%	72.00%
3	30	32	16	18	19	34	149
	60%	64%	32%	36%	38%	68%	49.67%
4	28	19	9	20	19	36	131
	56%	38%	18%	40%	38%	72%	43.67%
5	23	26	24	30	4	42	149
	46%	52%	48%	60%	8%	84%	49.67%

Conjunto de 50 imágenes por clase 64x64 px a color, 6 características en cada uno de los niveles.

Cuadro 4.16: WDM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, 5 niveles, 6 características por nivel).

	ÁRBOL DE CLASIFICACIÓN						
Nivel	MG	MF	CE	HO	NU	CI	Total Imágenes
1-5	48	49	43	47	48	46	281
	44%	98%	86%	94%	96%	92%	93.67%

Conjunto de 50 imágenes por clase 64x64 px a color, 30 características (6 por nivel).

Cuadro 4.17: WDM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (300 imágenes, nivel 1-5, 32 características).

El cuadro 4.17 muestra en número y porcentaje las imágenes clasificadas utilizando 30 características (6 por nivel), excluyendo la desviación estándar y media ambas del componente de aproximación. A lo cual para el Árbol de Clasificación y Regresión las características en conjunto le permite clasificar el 93.67 % de las 300 imágenes.

MATRICES DE CONFUSIÓN

Se reporta la matriz de confusión para detectar el número de imágenes clasificadas erróneamente por el Árbol de Clasificación, tanto para la matriz de co-ocurrencia de grises como para la matriz wavelet bajo el mismo conjunto de prueba, donde se aprecia que las características que mejor describen a las imágenes pertenecen a la wavelet por su menor número de incidencias erróneas.

Matriz de Confusión: Matriz de Coocurrencia de Niveles de gris (50 imágenes por clase). Método: Arbol de Decisión.

	MG	MF	CE	HO	NU	CI
MG	100	0	0	0	0	0
MF	0	97	3	0	0	0
CE	0	0	93	7	0	0
HO	0	1	2	97	0	0
NU	0	0	0	1	99	0
CI	0	0	0	5	1	94

Matriz de Confusión: Matriz Wavelet 2D (50 imágenes por clase). Método: Arbol de Decisión.

	MG	MF	CE	HO	NU	CI
MG	100	0	0	0	0	0
MF	0	98	2	0	0	0
CE	0	0	100	0	0	0
HO	0	0	0	100	0	0
NU	0	0	2	1	97	0
CI	0	0	0	0	0	100

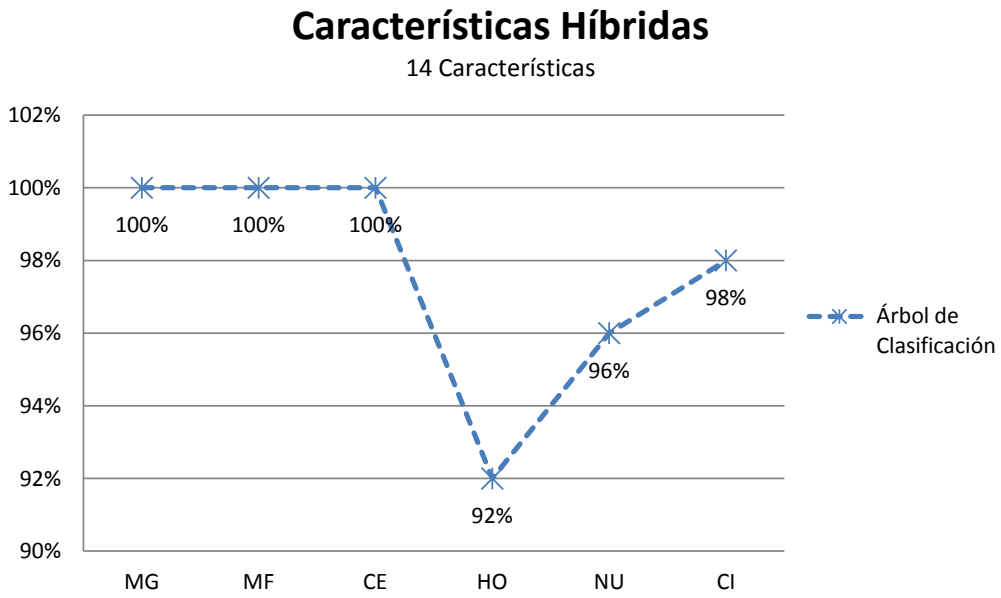
4.3.3. Características Híbridas

Una vez contando con características tanto de la GLCM como la WDM, formaremos un nuevo primer conjunto de 14 características el cual está compuesto por:

- 6 características de la GLCM:
  - 5 características propias de la GLCM (C,H,M,N,V)
  - Desviación estándar de la imagen (que en pruebas anteriores demostró aportar información importante para la clasificación)
- 8 características de la WDM que son :
  - Media y Desviación estándar (ambas del componente de aproximación)
  - 6 características del segundo nivel (las cuales en pruebas anteriores resultaron características favorables para la clasificación)

El cuadro 4.18 muestra el análisis en el Árbol de Clasificación, con base en el conjunto de las 14 características anteriormente descritas, extraídas de un conjunto de 50 imágenes de 64x64 pixeles. En el cual clasifica con un 97.67 % donde las clases que clasifica al 100 % son: MG, MF y CE, CI con 98 %, NU con 96 % y HO com 92 % (véase cuadro 4.19).

La Tabla 8.1 describe en número y porcentaje las imágenes clasificadas exitosamente sobre el conjunto descrito anteriormente.



Cuadro 4.18: WDM+GLCM. Gráfica de imágenes clasificadas por el Árbol de Clasificación y Regresión (12 características, 300 imágenes).

ÁRBOL DE CLASIFICACIÓN						
MG	MF	CE	HO	NU	CI	Total Imágenes
50	50	50	46	48	49	293
100%	100%	100%	92%	96%	98%	97.67%

6 Características GLCM + 8 Características de WDM
(Media y D.E del componente de aproximación+ 6 características del 2° nivel).

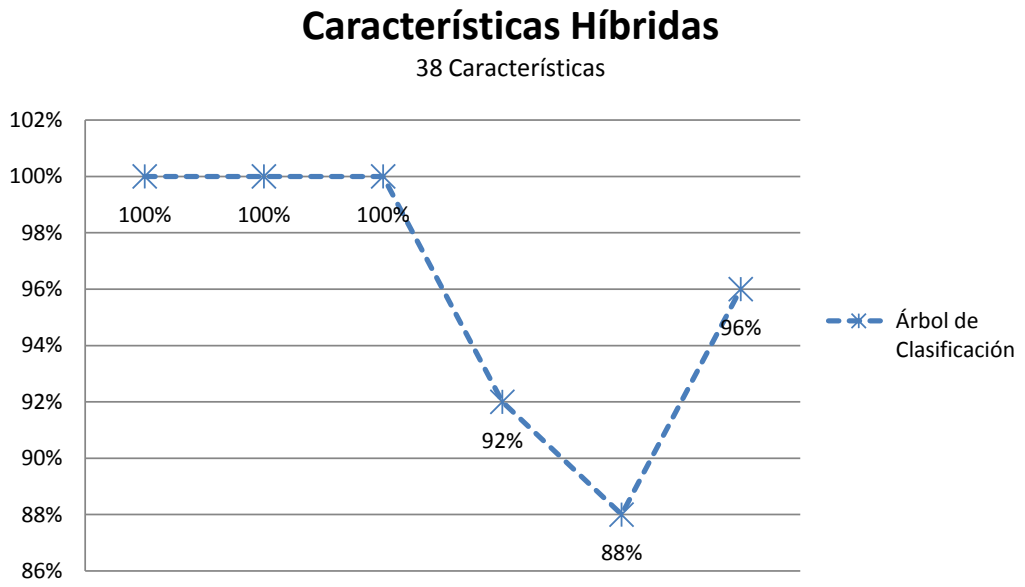
Cuadro 4.19: WDM+GLCM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (12 características, 300 imágenes)

Ahora se forma el segundo nuevo conjunto de 38 características:

- 6 características de la GLCM:
  - 5 características propias de la GLCM (C,H,M,N,V)
  - Desviación estándar de la imagen (que en pruebas anteriores demostró aportar información importante para la clasificación)
- 32 características del vector de la WDM (todas las características de todos los niveles).

El cuadro 4.20 muestra el analisis del Árbol de Clasificación y Regresión con base en el conjunto de datos anteriormente descrito donde se observa que clasifica con un 96 % las clases de las cuales: MG, MF y CE clasifico al 100 %.

El cuadro 4.21 describe el Árbol de Clasificación y Regresión en número y porcentaje de imágenes clasificadas exitosamente sobre el conjunto descrito anteriormente.



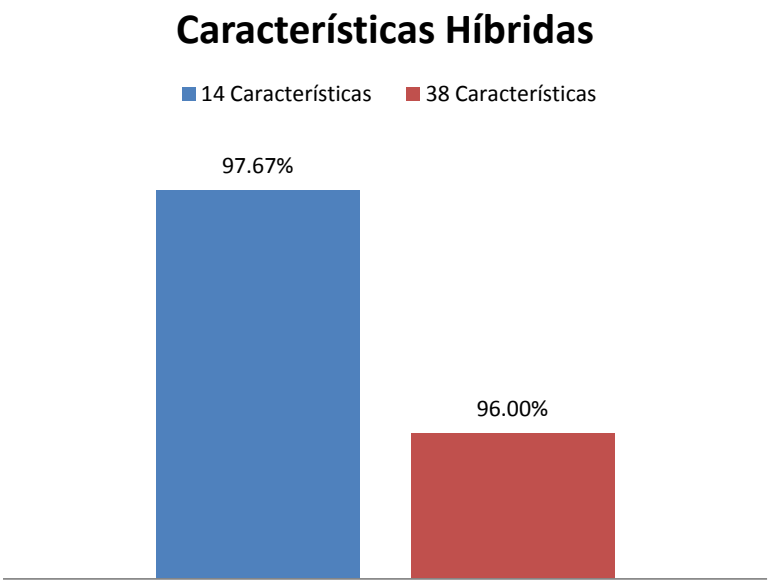
Cuadro 4.20: WDM+GLCM.Gráfica de imágenes clasificadas por el Árbol de Clasificación y Regresión (38 características, 300 imágenes)

ÁRBOL DE CLASIFICACIÓN						
MG	MF	CE	HO	NU	CI	Total Imágenes
50	50	50	46	44	48	288
100%	100%	100%	92%	88%	96%	96.00%

6 Características GLCM + 32 Características WDM
(Todas las características de todos los niveles).

Cuadro 4.21: WDM+GLCM. Tabla de imágenes clasificadas por el Árbol de Clasificación y Regresión (38 características, 300 imágenes)

El cuadro 4.22 muestra la comparación de características híbridas, tanto de las 14 como las 38 características propuestas anteriormente. En el cual el uso de las 14 características obtuvo su mejor desempeño con el 97.67 % a comparación de las 38 caractesísticas con un 96 %.



Gráfica 10

Cuadro 4.22: WDM+GLCM. Gráfica comparativa de caracterísitcas híbridas: 14 y 38 características.

El mejor conjunto de características híbridas la proporciona el conjunto de 14 con una diferencia un tanto significativa de 1.67 %.

Capítulo 5

Conclusiones y Trabajo Futuro

En este trabajo evaluamos dos métodos de extracción de características de textura: la Matriz de Co-ocurrencia de Niveles de Gris y la Transformada Wavelet Haar 2D para la clasificación de células con la finalidad de identificar patrones de tinción asociados a la fase celular.

Después de evaluar ambos métodos en una base de imágenes con 720 células los resultados obtenidos suguieren que las características basas en la transformada discreta Wavelet tienen mejor desempeño que los descriptores de textura obtenidos con las propiedades estadísticas de primer orden como la energía, el contraste, la media y la varianza. Así, mediante el uso de la transformada wavelet es posible clasificar el 98 % de las imágenes con células cuando se evalua el algoritmo desarrollado mediante el método de dejar uno afuera. El uso de características híbridas no es suficiente para superar a la transformada wavelet.

Como trabajo futuro se propone utilizar los 16 momentos estadísticos a partir de la matríz de co-ocurrencia y evaluar algunos otros métodos de transformada wavelet más recientes. Tambien sería interesante evaluar el agregar propiedades geométricas de las células aplicando otras técnicas del procesamiento digital de imágenes. Finalmente en cuanto a los algoritmos de aprendizaje automático se pueden evaluar diferentes métodos de clasificación como las redes neuronales.

Bibliografía

[1] H.Muller. *A Review of Content-Based Image Retrieval Systems in Medical Applications: Clinical Benefits and Future Directions. International Journal of Medical Informatics*, **2004**, *Vol. 73*, 1.

[2] D.W.Bates. *The Costs of Adverse Drug Events in Hospitalized Patients. Journal of the American Medical Association*, **1997**, *Vol. 277, no. 4*, 307.

[3] N.S.Weingar. *Epidemilogy of Medical Error. British Medical Journal*, **2000**, *Vol. 320*, 774.

[4] D.Copec. *Human Errors in Medical Practice: Systematic Classification and Reduction with Automated Information Systems. Journal of Medical Systems*, **2003**, *Vol. 27*, 297.

[5] Chen.X; Zhou.X; Wong.STC. *Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. IEEE Transactions on Biomedical Engineering*, **2006**, *Vol. 53*, 762.

[6] Neumann.B; Walter.T; Hériché.JK; Bulkescher.J; Erfle.H; Conrad.C; Rogers.P; Poser.I; Held.M; Liebel.U; Cetin.C; Sieckmann.F; Pau.G; Kabbe.R; Wünsche.A; Satagopam.V; Schmitz.MHA; Chapuis.C; Gerlich.DW; Schneider.R; Eils.R; Huber.W; Peters.JM; Hyman.AA; Durbin.R; Pepperkok.R; Ellenberg.J. *Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature*, **2010**, p 721.

[7] Lu.J; Liu.T; Yang.J. *Automated cell phase classification for zebrafish fluorescence microscope images. 20th International Conference on Pattern Recognition*, **2010**.

[8] Harder.N; Mora-Bermúdez.F; Godinez.WJ; Ellenberg.J; Eils.R; Rohr.K. *Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. Med Image Comput Assist*, **2006**, *Vol. 9*, 840.

[9] Ranjan.Parekh. *Using textture analysis for medical diagnosis,Jadavpur University,India. IEEE Computer Society*, **2012**, p 28.

[10] R.Wang; A.R.Hanson; E.M.Riseman. *Texture Analysis Based on Local Standard Deviation of Intensity. Conf. Computer Vision and Pattern Recognition, IEEE CS Press*, **1986**, p 482.

[11] R.M.Haralick. *"Statistical and Structural Approaches to Texture" , Proc. IEEE*, **1979**, *Vol. 67*, 786.

[12] A.Haar. *On the Theory of Orthogonal Function Systems, Mathematische Annalen*, **1910**, *Vol. 69*, 331.