UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Programa de Maestría y Doctorado en Ingeniería
Ingeniería Eléctrica – Procesamiento Digital de Señales

Segmentación semi-automática de células hipofisarias de ratón con interfaz gráfica.

TESIS

QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN INGENIERÍA

PRESENTA:
ERIKA ARACELI GONZÁLEZ VILLA

Directores de Tesis

Dra. Lucía Medina Gómez
Facultad de Ciencias

Dr. Mathieu Hautefeuille
Facultad de Ciencias

Ciudad Universitaria, Cd. Mx. Julio 2017

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Contenido

Resumen

La palabra célula fue acuñada por primera vez por el científico Robert Hooke en 1665 en su libro *Micrographia* en el que describía sus observaciones de finas rebanadas de corcho mediante un microscopio. Denominó como células o poros a las formaciones que veía aunque no sabía que estaba observando las paredes celulares de la planta con la que elaboraban el corcho.

Mucho tiempo ha pasado y actualmente el estudio de las células ha mejorado bastante gracias al desarrollo tecnológico del microscopio, por lo que el problema se ha cambiado de obtener imágenes de células al adecuado procesamiento de la enorme cantidad de información que generan los equipos que actualmente se usan para la generación de imágenes celulares.

En especial, la mayoría de las veces, para poder implementar un análisis de las células a través de imágenes se requiere como primer paso su identificación, y para hacerlo de manera automatizada mediante la computadora se suelen utilizar técnicas de procesamiento digital de imágenes que en su mayoría requieren la detección de los bordes de las células. Una vez que se han identificado las células se suele trabajar en diferentes tipos de análisis sobre las mismas. Hay que resaltar que las técnicas de segmentación de células suelen ser diferentes para cada tipo de imagen y es casi imposible generar un algoritmo universal, es decir, uno que funcione para todo el abanico de posibles imágenes.

El presente trabajo busca contribuir con una técnica para la detección de células hipofisarias en stacks de imágenes adquiridas mediante microscopía de fluorescencia. Ya que las células en este tipo de imágenes no tienen bordes visibles no es posible utilizar alguna de las técnicas tradicionales, sin embargo, se tomó como base las características que presentan las células, específicamente el comportamiento de su serie de tiempo de la intensidad, con esto se realizó un análisis de 5 características de las series de tiempo, área bajo la curva, promedio, desviación estándar, curtosis y sesgo, de este resultó que sólo tres de ellas son relevantes y con ellas se implementaron dos clasificadores de regresión logística, uno que usa el área bajo la curva y el sesgo mientras que el otro usa la curtosis y el sesgo.

Antes de poder aplicar el clasificador se hizo una corrección sobre los stacks de imágenes ya que debido a la técnica que se usó para su adquisición (microscopía de fluorescencia) dichos stacks presentan fotoblanqueamiento, que es el oscurecimiento de las imágenes en el tiempo, entonces para que la tarea de comparar las series de tiempo de las células fuera más sencilla se buscó corregir los stacks mediante una técnica de procesamiento digital de imágenes llamada especificación del histograma.

Una vez corregidos se eligieron 120 series de tiempo, 60 asociadas a células y 60 asociadas a ruido provenientes de 6 diferentes stacks, las cuales fueron utilizadas para el entrenamiento de los dos clasificadores. Posteriormente se usaron otras 60 series de tiempo diferentes, 30 ligadas a células y 30 a ruido, con esto se asoció un error a los clasificadores que resultó de 17% para el clasificador Área-Sesgo y 15% para el de Curtosis-Sesgo.

Finalmente ambos clasificadores fueron puestos a prueba con 12 stacks de imágenes para conocer su desempeño, para esto se consideró que actualmente la identificación de las células se realiza de manera manual y requiere que el experto destine una enorme cantidad de tiempo para dicha tarea, por lo que se buscó que este tiempo se viera reducido. Teniendo esto en mente se pudo observar que el clasificador curtosis-sesgo se desempeña de forma similar a como lo hace el experto al seleccionar las células manualmente en un 91% de las veces, a diferencia del 71% del clasificador área-sesgo. Este porcentaje fue tomado sobre el número de células identificadas por el clasificador y no sobre el número de células que podría seleccionar el experto.

Con esto se pudo observar que la identificación de las células mediante la clasificación de sus series de tiempo asociadas dio resultados satisfactorios, sobre todo pensando que las características usadas fueron bastante sencillas, por lo que se espera que al añadir nuevas características al clasificador este pueda incrementar el número de células identificadas en los stacks, y que de esta forma este algoritmo se pueda implementar como una técnica más para la segmentación de células sobre imágenes de la hipófisis que contiene un gran número de ellas y que no presentan bordes distinguibles.

Introducción

Según un estudio en psicolingüística del Instituto Max Planck [1], el sentido de la vista es al que los seres humanos le dan más importancia por encima del resto de los sentidos, esto podría deberse a que la mitad del cerebro humano se dedica a tareas relacionadas con la visión [2]. No obstante la tarea de procesar esta información no es sencilla, a diferencia del cerebro humano que es capaz de reconocer y clasificar objetos de entre decenas de miles de posibilidades en fracciones de segundo [3], las computadoras actuales aún no son aptas para desempeñar adecuadamente este trabajo aunque su capacidad de realizar cálculos de forma rápida se ha incrementado de manera sustancial durante los últimos años, por esto último, áreas como el procesamiento digital de imágenes y la visión computacional, entre otras, han tenido un desarrollo importante en poco tiempo a la par del avance en los sistemas computacionales.

Ya que el sentido de la vista es uno de los más relevantes para el ser humano no es de extrañar que muchos de los desarrollos tecnológicos estén enfocados al análisis visual del mundo que nos rodea, iniciando con el telescopio y el microscopio óptico, pasando por la cámara fotográfica y llegando a sus contrapartes electrónicas y digitales, además hay que incluir los dispositivos que nos permiten "ver" a través de los entes como las máquinas de ultrasonido y rayos X, o los equipos de tomografía por emisión de positrones (PET).

Uno de los dispositivos mencionados que han marcado un interés particular en el campo de la biología y en las ciencias de la salud es el microscopio, desde su invención por Zacharias Jansen en 1950 (aunque algunos lo atribuyen a Galileo Galilei [4]) ha revolucionado dichos campos, en particular la biología celular encargada del estudio de la fisiología de las células y las interacciones entre ellas así como con su medio. Con esto se ha permitido la comprensión y estudio de enfermedades asociadas a su mal funcionamiento como el cáncer o el Alzheimer, esto ha sido posible gracias al desarrollo tecnológico del microscopio que ha evolucionado de tal forma que ahora permite la adquisición de una gran cantidad de imágenes en poco tiempo.

En consecuencia a esto último, ha sido necesaria la implementación de tecnologías computacionales que permitan el análisis en poco tiempo de la gran cantidad de información que se genera, sobre todo en el área de la biología celular donde uno de los problemas más habituales es la identificación de células en las imágenes y ,para esto, se basan en el procesamiento digital de imágenes, que es un conjunto de técnicas que buscan realzar y restaurar imágenes para dar una mejor interpretación de ellas, además de permitir segmentar y describir los elementos presentes [5].

El caso de interés en este trabajo recae en la detección de células hipofisarias de ratones en stacks de imágenes adquiridas mediante microscopía de fluorescencia, a pesar de que existen muchísimos trabajos alrededor de la identificación automatizada de células en imágenes implementado diferentes métodos, en su mayoría tienen ciertas características que son similares entre sí:

- Las imágenes contienen un número reducido de células ya que muchas veces buscan identificar partes internas de las mismas, como el núcleo, el citoplasma, etc. [6]
- Si se tiene un número reducido de células en la imagen es fácil identificar “a ojo” los bordes de las células, y de esta manera se pueden aplicar diferentes técnicas de procesamiento digital de imágenes para la detección de bordes como los filtros gaussianos o laplacianos [7].
- A veces las células no están dentro de una imagen sino dentro de un stack de imágenes, donde cada imagen se refiere a un punto en el tiempo, y dentro de estas imágenes se tienen células en movimiento, por lo que se desea poder cuantificar el movimiento de las células y para esto se aplican técnicas de tracking [8], usualmente esto es posterior a la segmentación de las células.
- En otras ocasiones la morfología de las células es muy complicada o pueden existir casos de células aglomeradas, por lo que el proceso de segmentación mediante filtros ya no resulta tan sencillo [9].

Debido a estas características, los métodos empleados comúnmente en la segmentación de células son los siguientes [10]:

Umbral de intensidad: es la más utilizada y se basa en que las células poseen una intensidad diferente a la del fondo.
Detección de características: como sus bordes o si presentan alguna forma relativamente invariante se pueden usar filtros de detección de bordes.
Filtros morfológicos en escala de grises y binarios: en realidad son una forma de pre o pos-procesamiento a la segmentación respectivamente, el primero para mejorar estructuras en la imagen antes de la segmentación y el segundo para mejorar zonas ya segmentadas.
Acumulación de regiones: en este caso se utilizan puntos semilla previamente seleccionados para conectar de forma iterativa puntos y formar regiones, por ejemplo el crecimiento de regiones, otro ejemplo es la transformación “watershed” [11] que se basa en la morfología matemática de la imagen para implementar la segmentación.
Modelos deformables: requiere una segmentación burda previa y busca ajustar un contorno, superficie, etc. al objeto buscado de manera iterativa.

Como se puede observar en la mayoría de estos casos se requiere el uso de los bordes de las células para su segmentación, sin embargo existen ocasiones en el que se desea encontrar células dentro de imágenes donde hay una gran cantidad de ellas y no es posible identificar bordes, como es el caso de el presente trabajo, por lo que es necesaria la implementación de nuevas técnicas de segmentación de células que permitan resolver esta problemática y así contribuir al avance en el estudio de la biología celular y las disciplinas satélite.

Objetivos

El propósito de este trabajo es implementar la detección semi-automatizada de células hipofisarias en imágenes adquiridas mediante microscopía de fluorescencia, tomando en cuenta las series de tiempo de la intensidad asociadas a las células, ya que no es posible hacer uso de sus bordes dado que no son visibles por la enorme cantidad de células presentes en el tejido.

El objetivo principal es que dicha detección mejore el tiempo requerido por un experto para la identificación manual de las células.

Se implementarán dos clasificadores de regresión logística basados en en una pareja de características diferente y se compararán los resultados. La idea es elegir el que presente el mejor funcionamiento e implementarlo en el futuro dentro de una interfaz gráfica.

1. Las células hipofisarias

“Every single day your body produces more new cells than there are atoms in the universe”
Michio Kaku

Los primeros estudios relacionados con las hormonas indicaban que debería existir alguna comunicación química entre los diferentes órganos de un animal, también se habían registrado casos donde personas recibían tratamientos exitosos cuando se les administraba extractos de tejido endocrino de algún animal [12], pero no fue hasta junio de 1905 cuando Ernest Starling acuñó la palabra hormona como “*the chemical messengers which speeding from cell to cell along the blood stream, may coordinate the activities and growth of different parts of the body*” [13], es decir los mensajeros químicos que van de célula en célula a través del flujo sanguíneo y que pueden coordinar las actividades y crecimiento de diferentes partes del organismo.

A partir de esto, muchos científicos volcaron su interés por estos mensajeros químicos, y con esto se descubrió que muchas especies presentan glándulas endocrinas que secretan hormonas similares entre sí, aunque sus efectos pueden ser diferentes [14], posteriormente entre 1940 y 1955 Harris y Green establecieron la existencia de una conexión entre el sistema nervioso y el sistema endocrino en la que el primero dirige a la glándula pituitaria para la correcta segregación de hormonas además de mostrar la conexión vascular entre esta glándula y el hipotálamo [15]. Actualmente se sabe que la secreción hormonal está influenciada, entre otros, por el sistema nervioso central mientras que el cerebro se ve influenciado por la acción de las hormonas [16].

Además de que el hipotálamo y la pituitaria permiten un vínculo entre los sistemas nervioso y endocrino, la glándula pituitaria juega un papel trascendente en el organismo ya que segrega hormonas para estimular o inhibir la secreción de hormonas de otras glándulas endocrinas, razón por la que se le conoce también como glándula maestra. Dicha glándula posee diferentes tipos de células que son las encargadas de la segregación de hormonas y estudios recientes demuestran que estas células están interconectadas formando redes tridimensionales para la correcta segregación de hormonas [17], estas redes guardan cierta analogía con las formadas por las células neuronales que han sido ampliamente estudiadas y que incluso permiten el estudio de enfermedades [18].

Ya que este trabajo es parte de un proyecto más amplio donde se busca desarrollar análisis de redes de células hipofisarias de manera semi-automática que permitan también el análisis de enfermedades relacionadas con la secreción de hormonas, en este capítulo se describirá con más detalle el ambiente en el que se encuentran dichas las células y cómo mediante una técnica de imagenología es posible realizar un análisis de ellas.

1.1 Las hormonas

Las señales químicas o ligandos son moléculas liberadas en alguna posición y que viajan a otra ubicación para producir alguna respuesta, se pueden dividir en señales intracelulares e intercelulares, las primeras son moléculas que se liberan dentro de la misma célula y viajan a través de ella para adherirse a algún receptor, las segundas son moléculas producidas en alguna parte de una célula y liberadas en el fluido intersticial[1], posteriormente se unen a otra célula que posea el receptor adecuado para la molécula. Los receptores son proteínas o glicoproteínas cuyas características químicas o morfológicas hacen que solo un tipo de señal química se una a ellos, a esto se le conoce como especificidad [19].

Específicamente las hormonas son señales químicas intercelulares que después de haber sido liberadas en el fluido intersticial pasan al torrente sanguíneo y se enlazan a las células de algún tejido, el resultado es influir en alguna actividad específica del órgano en cuestión.

La secreción de hormonas se regula de una, dos o tres de las siguientes maneras dependiendo del tipo de hormona:

- Nivel de algún químico en la sangre, por ejemplo la hormona insulina es regulada por el nivel de azúcar en la sangre.
- La secreción de algunas hormonas está regulado por otras hormonas, por ejemplo la hormona estimulante de la tiroides que como su nombre lo indica estimula la glándula tiroides para que segregue hormonas.
- El sistema nervioso, por ejemplo la hormona epinefrina (adrenalina) es liberada por el sistema nervioso.

1.1.2 El sistema endocrino

Las hormonas se encargan de diferentes actividades metabólicas y fisiológicas en el organismo entre las que se encuentran la regulación del crecimiento, desarrollo y procesos reproductivos, la estimulación de la actividad de algunos músculos y glándulas, la participación dentro del ritmo circadiano, entre otras. El sistema endocrino, que es un conjunto de glándulas y agrupaciones de células epiteliales[2] especializadas es el responsable de su secreción.

Los órganos que poseen células endocrinas son el hipotálamo, el timo, el páncreas, los ovarios, los testículos, los riñones, el estómago, el hígado, el intestino delgado, la piel, el corazón, el tejido adiposo y la placenta. Por otra parte las glándulas endocrinas incluyen a la hipófisis, tiroides, paratiroides y las glándulas pineal y suprarrenales, ver Figura (1.1).

[1] Líquido que baña y se encuentra en el espacio entre las células, está conformado por agua, sales, ácidos grasos, aminoácidos, productos de desecho de las células, etc.
[2] Son células del tejido que recubre superficies del cuerpo tanto externas como internas.

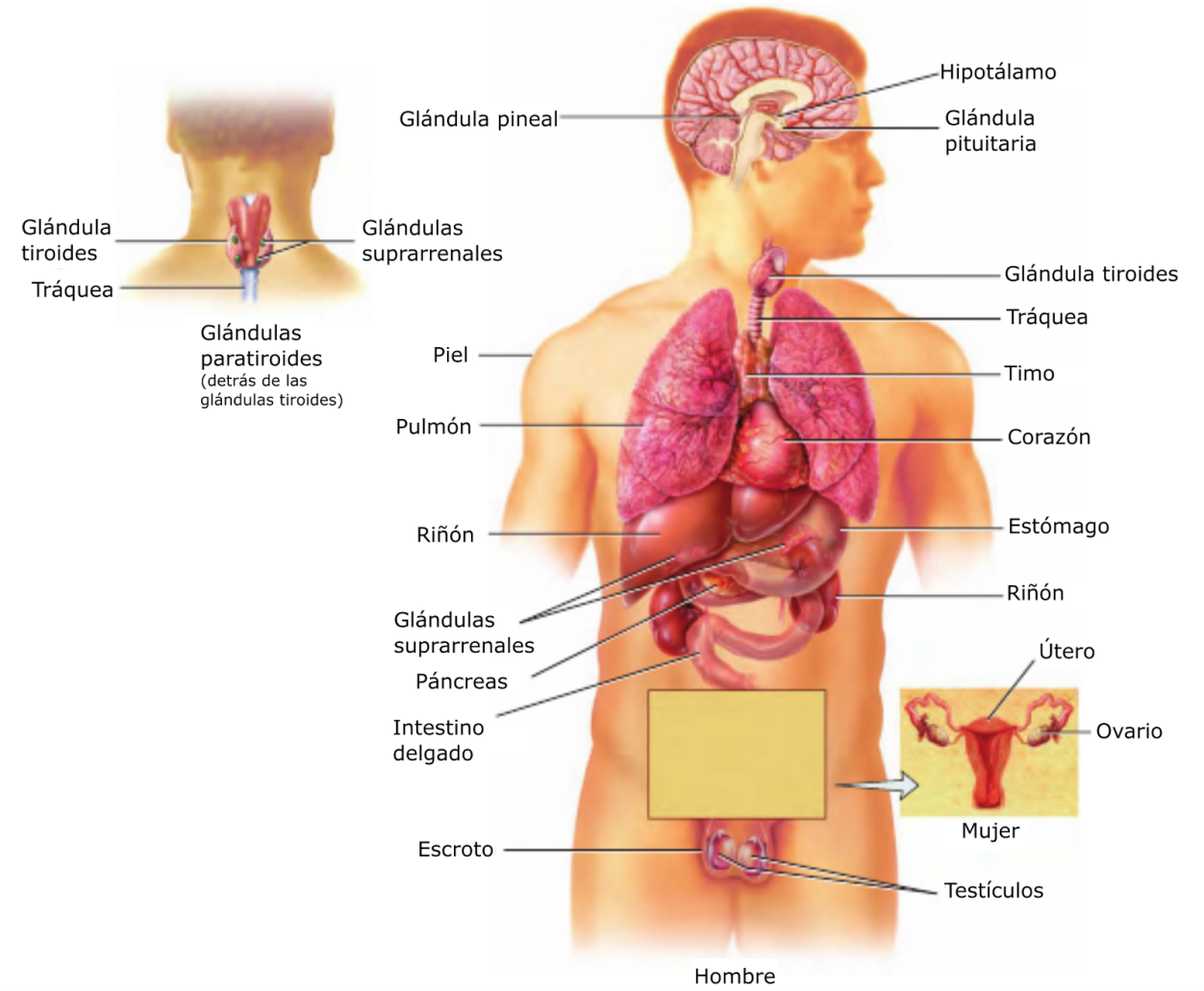

Figura 1.1 Glándulas endocrinas y órganos que poseen células endocrinas. Imagen obtenida de [20]

Cabe señalar que además de las glándulas endocrinas existen las glándulas exocrinas, las cuales, a diferencia de las primeras, secretan sus productos en ductos que los transportan hacia alguna cavidad o a la superficie, como las glándulas sudoríparas o las glándulas salivales, mientras que las glándulas endocrinas secretan las hormonas en el flujo sanguíneo.

1.2 La hipófisis

Como se ha mencionado antes, la hipófisis, o pituitaria, es una glándula endocrina que se ubica en una concavidad del hueso esfenoides[3], en la base del cerebro [21], ver Figura (1.2), la pituitaria secreta hormonas que a su vez controlan la liberación de hormonas en otras glándulas endocrinas. A grosso modo está formada por los lóbulos anterior y posterior, el primero secreta las siguientes hormonas:

[3] Hueso situado en la base del cráneo

- TSH (tirotropina): hormona estimulante de tiroides
- ACTH (adrenocorticotrópica): estimula las glándulas suprarrenales
- LH (luteinizante): estimula la ovulación o la producción de testosterona en mujeres y hombres respectivamente.
- FSH (estimulante de los folículos): regula el desarrollo, crecimiento y maduración en la pubertad, y los procesos reproductivos.
- Prolactina: estimula la producción de leche en las glándulas mamarias
- Somatotropina: hormona del crecimiento

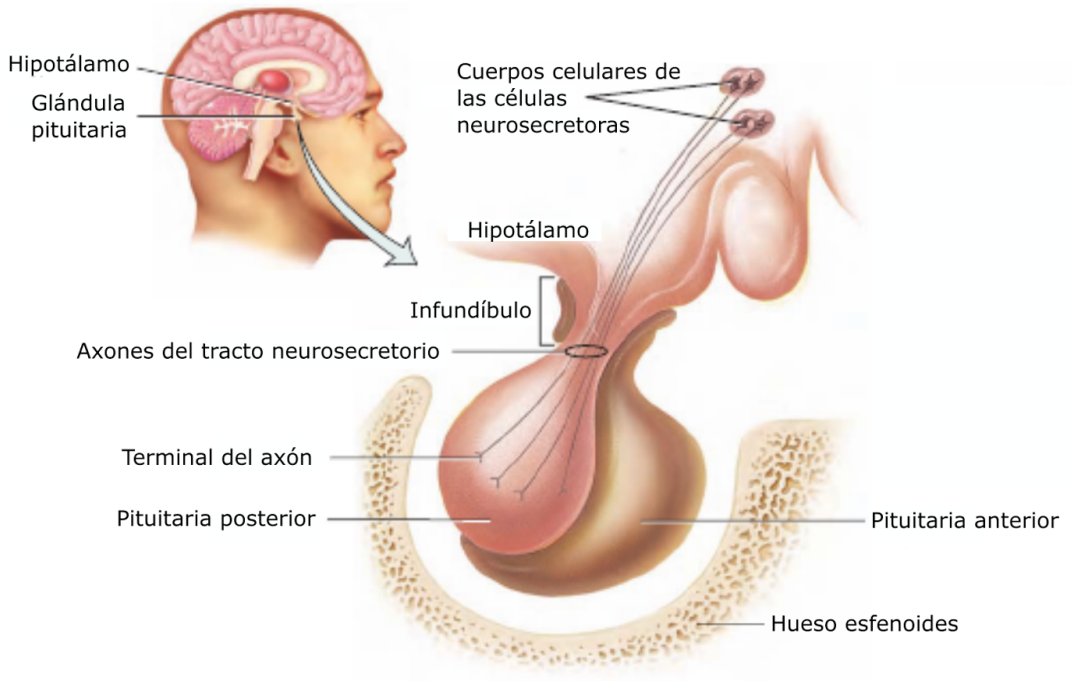

Figura 1.2 Glándula hipófisis y sus partes. Imagen obtenida de [20].

Por otra parte el lóbulo posterior almacena y libera las hormonas oxitocina (estimula las contracciones del útero en el parto y la salida de leche durante la lactancia) y vasopresina (estimula la retención de agua por los riñones), además está conectado al hipotálamo que es una región del cerebro que permite la conexión entre el sistema nervioso y el sistema endocrino, como se ha mencionado antes el sistema nervioso es capaz de estimular o inhibir la liberación de hormonas por parte del sistema endocrino. La localización y partes de la pituitaria se muestran en la Figura (1.2).

En cierto sentido el sistema nervioso y el sistema endocrino presentan ciertas similitudes, ambos utilizan ligandos y receptores para comunicarse entre células, sin embargo mientras el sistema nervioso se basa en el uso de conexiones entre las células (neuronas) para la transmisión de la información el sistema endocrino utiliza como medio el torrente sanguíneo, por esta razón en el sistema nervioso la comunicación es casi inmediata, en cuestión de segundos, mientras que el sistema endocrino requiere de una cierta cantidad de tiempo para que se presente, que puede ir desde minutos hasta horas [22]. Además los efectos debidos a la comunicación en el sistema nervioso se presentan a corto plazo y en un

tejido u órgano específico mientras que en el sistema endocrino dichos efectos son más duraderos y suelen tener un efecto más general en el organismo [19].

1.2.1 Células hipofisarias

Otra característica que pone de manifiesto la diferencia entre el sistema nervioso y el sistema endocrino es la cantidad de tipos de células dedicadas a la comunicación entre células, mientras el primero tiene 2, las neuronas y las células neurogliales, el segundo presenta una enorme variedad de tipos de células, tan solo en el lóbulo anterior de la hipófisis existen 5 tipos de células secretoras de hormonas, mostradas en la Tabla (1.1).

Tipo celular	Hormona que secreta	% del total de células secretoras	Localización dentro del lóbulo anterior de la hipófisis.
Somatotropas	Hormona del crecimiento	50	Zonas laterales
Lactotropas	Prolactina	10-30	Distribución aleatoria
Corticotropas	Hormona ACTH	10	Zona anterior y media
Tirotropas	Hormona TSH	5	Zona anterior, media y lateral
Gonadotropas	Hormonas FSH y LH	20	Todo el lóbulo

Tabla 1.1 Células que segregan hormonas en el lóbulo anterior de la hipófisis. Obtenida de [21], [22]

Debido a que anteriormente no se contaba con las herramientas adecuadas para el estudio de células dentro de los tejidos lo que se hacía en general era tomar una rebanada muy delgada de la pituitaria de algún animal, usualmente ratones, y realizar un análisis de las células en el tejido mediante microscopio. Con los primeros estudios se reveló que las células en la hipófisis estaban distribuidas de manera homogénea, además esto provocó el análisis del funcionamiento de las células fuera de su tejido original concluyendo que las células liberan pulsos hormonales al flujo sanguíneo como respuesta a las señales proporcionadas por el hipotálamo sin presentar interacciones entre ellas [23].

Sin embargo con el mejoramiento de las técnicas de imagenología ha sido posible estudiar de mejor forma los tejidos, en particular la hipófisis donde se ha descubierto que en realidad las células están organizadas en redes tridimensionales que facilitan la emisión de los pulsos hormonales mediante la propagación de señales entre células vecinas y distantes a través de la misma glándula [24]. La importancia de las redes tridimensionales se ha comprobado al estudiar monocapas de células endocrinas ya que los resultados muestran una disminución en la magnitud de hormona liberada, demostrándose así que las células requieren de un funcionamiento como población, de hecho actualmente se sabe que la secreción de pulsos hormonales hacia el torrente sanguíneo requiere de la actividad conjunta de cientos de miles de células endocrinas [25].

El estudio de dichas redes en la hipófisis es bastante reciente [23], [24], sin embargo en el caso del sistema nervioso ya se tienen bastantes avances, específicamente en el análisis de las redes de neuronas [26], [27]. Incluso se han desarrollado plataformas que buscan realizar el análisis de estas redes neuronales a partir de videos que muestran la actividad de las neuronas en el tejido [28]. Sin embargo tanto en los estudios de redes de neuronas y de redes de células de la hipófisis se analiza un número reducido de células debido a que las técnicas de imagenología a pesar de haber mejorado enormemente aún no permiten ver una gran porción del tejido o una profundidad adecuada o los equipos son demasiado costosos, otro problema es que la detección de las células en los videos puede requerir demasiado tiempo y la automatización de la detección de las células puede ser complicada si los videos no presentan la mejor resolución.

En la siguiente sección se describe la técnica conocida como microscopía de fluorescencia que permite la adquisición de imágenes de células hipofisarias, es importante conocer este método ya que este es el tipo de imágenes que se utilizaron durante el desarrollo de este trabajo.

1.3 Imágenes de células hipofisarias mediante microscopía de fluorescencia

Una de las técnicas empleadas en el campo de la biología para el estudio de diferentes tejidos, células y estructuras celulares a través de imágenes es la microscopía de fluorescencia que se basa en el fenómeno por el cual al hacer incidir luz de una longitud de onda determinada sobre cierto tipo de sustancias ellas son capaces de emitir luz con una longitud de onda mayor que la de incidencia después de haber transcurrido algunos nanosegundos [29], a este comportamiento se le denomina fluorescencia.

Físicamente lo que ocurre con estas sustancias es que contienen moléculas denominadas fluoróforos las cuales poseen grupos funcionales (esencialmente conjuntos específicos de átomos) que se encuentran en su estado base antes de hacer incidir luz en ellos, en el instante en que se iluminan con una longitud de onda específica los electrones de las capas externas del grupo funcional absorben la energía debida a lo fotones y pasan a un estado excitado, pero no permanecen ahí por mucho tiempo regresando al estado base y al mismo tiempo emitiendo un fotón con energía menor a la del fotón absorbido.

Estos fotones emitidos son los que se emplean para la adquisición de imágenes, regresando al caso de la biología se utiliza una luz de incidencia sobre el espécimen a estudiar que puede ser un láser, una lámpara de mercurio, etc., y se hace uso de sensores que detectan los fotones emitidos gracias a los fluoróforos filtrando además la luz de incidencia, posteriormente sigue el proceso de digitalización de la imagen así adquirida. Es claro que ante un mayor número de fotones detectados por el sensor se tendrá un nivel de intensidad mayor en el valor del píxel y viceversa. Cuando junto con esta técnica de adquisición de imágenes se hace uso además de un microscopio se le conoce como microscopía de fluorescencia.

Aunque existen muchas sustancias naturales que ya son fluorescentes por sí mismas, como la clorofila de las plantas o la proteína verde fluorescente (GFP) producida por la medusa aequorea victoria, se han

sintetizado en laboratorio fluoróforos que pueden ser utilizados para diferentes fines mediante la microscopía de fluorescencia. Estos fluoróforos son introducidos en los ejemplares a estudiar y permiten ver las estructuras a través de la distribución de fluorescencia, también existen los denominados sensores fluorescentes [30] que se unen a ciertas moléculas o iones y aumentan o disminuyen su intensidad de luz de fluorescencia dependiendo de la concentración presente del elemento al que se ha unido, permitiendo realizar análisis en intervalos de tiempo sobre las muestras.

Regresando al caso de las imágenes de células de la hipófisis, se sabe que el requisito previo para que exista exocitosis[4] hormonal en las células de la pituitaria es la elevación de la concentración del calcio intracelular [Ca^{2+}] [31], es decir existe una correlación entre las señales de calcio intracelular y la cantidad de hormona segregada por las células de la hipófisis. El [Ca^{2+}] es un ión de calcio que permite la comunicación entre las células y que entra y sale de ellas a través de los canales de calcio[5]. Por otra parte existen fluoróforos que se adhieren al calcio intracelular, como el fluo-3 o el fluo-4, de esta forma se pueden medir indirectamente los pulsos hormonales al analizar el calcio intracelular mediante microscopía de fluorescencia.

1.3.1 Inconvenientes de la microscopía de fluorescencia

Al trabajar con microscopía de fluorescencia hay algunos problemas con los que es necesario enfrentarse, el principal es el que se refiere a la disminución en la intensidad de la luz de fluorescencia conforme transcurre el tiempo, conocido como fotoblanqueamiento, esto ocurre porque los fluoróforos no pueden realizar de manera ilimitada los cambios entre el estado base y el estado excitado, la causa es la producción de reacciones químicas entre el fluoróforo en su estado excitado y el oxígeno que modifican esta capacidad del fluoróforo mientras incide la luz [30], [32]; el tiempo de observación de una muestra se ve limitado por este obstáculo. La Figura (2.5) muestra el efecto del fotoblanqueamiento.

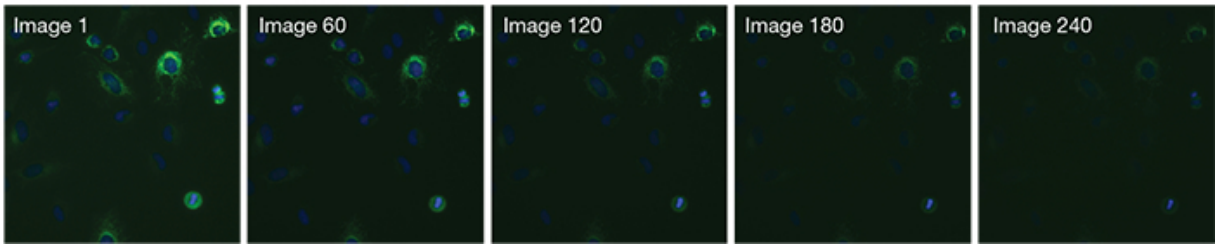

Figura 1.3 Imágenes de células HeLa adquiridas cada 15 s después de 2h de incubación, se puede observar cómo las imágenes se van oscureciendo con el tiempo, a este efecto se le conoce como fotoblanqueamiento. Imagen obtenida de [32]

Otros efectos debidos a la luz incidente es la fototoxicidad [33], que produce daños en las células, y la creación de sustancias químicas que la favorecen. Esto no es deseable cuando se busca analizar células vivas.

[4] Nombre formal del proceso por el que se secretan moléculas contenidas en vesículas dentro de una célula al espacio extracelular

[5] Canales en las membranas de las células que permiten a los iones de calcio pasar de un lado a otro.

Además de los efectos debidos a la iluminación, se presentan otros como la autofluorescencia, ver Figura (1.4), ya que muchos tejidos suelen emitir por naturaleza luz de fluorescencia con longitudes de onda similares a las que emiten los fluoróforos de uso común [34], esto se traduce en ruido en la imagen digital. Aunado a esto se presenta absorción y esparcimiento de la luz de fluorescencia, esto provoca una disminución en la intensidad de la luz que llega a los sensores. Asimismo si se desea trabajar con células vivas es importante mantenerlas en las condiciones adecuadas para su sobrevivencia, esto usualmente se logra mediante la perfusión del tejido que contiene las células y esto es causa de movimiento en las imágenes adquiridas que también se traduce en ruido.

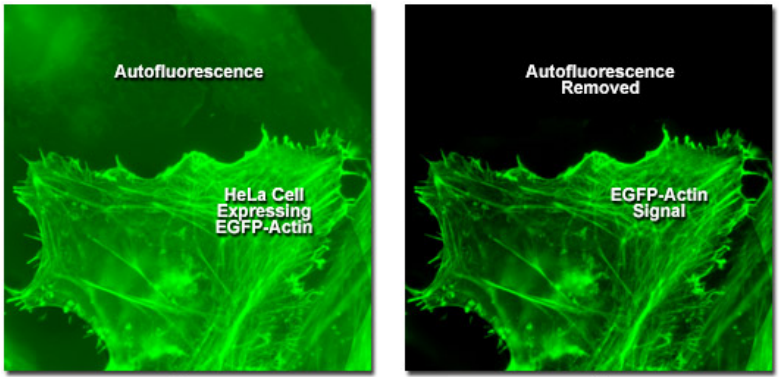

Figura 1.4 La imagen de la izquierda muestra una célula HeLa fluorescente con tejido circundante que también presenta fluorescencia, la imagen derecha muestra la misma célula con el tejido fluorescente removido. Imagen obtenida de [35].

El ruido de fondo es otra fuente de anomalías en las imágenes ocasionando que todas las mediciones de intensidad están dadas por el valor "real" de la intensidad + ruido de fondo [36], esto produce incertidumbre en las mediciones de intensidad. Existen varias fuentes de ruido de fondo, una de ellas es de la misma señal que se desea medir ya que el número de fotones que llegan al sensor producidos mediante fluorescencia tiene asociado un ruido de Poisson (shot noise) por tener un comportamiento estocástico descrito por la estadística de Poisson. Por otra parte el ruido de fondo puede provenir del mismo sensor ya que se producen de forma estocástica electrones térmicos en él debido a su calentamiento y no debido a la incidencia de fotones.

2. Clasificación de objetos

“Truth ... is much too complicated to allow anything but approximations”
John von Neumann

Ya que en este trabajo se ha usado un clasificador de regresión logística para la detección de células hipofisarias en stacks de imágenes de microscopía de fluorescencia en este capítulo se estudiará qué es un clasificador en general y en particular cómo funciona un clasificador de regresión logística.

2.1 El problema de la clasificación

La clasificación es al proceso de organizar objetos (observaciones, eventos, individuos) de acuerdo a sus propiedades (características, atributos, mediciones) mediante clases. Al realizar una clasificación los individuos de las mismas clases comparten las mismas propiedades por lo que se obtendrán diferentes clasificaciones dependiendo de las características que se elijan. Por ejemplo, si quisiéramos clasificar los animales de un acuario podríamos elegir como característica que sean peces y diferenciarlos de los que no lo son, incluso podríamos elegir más de una característica, por ejemplo diferenciar los peces amarillos del resto de los animales acuáticos.

Ahora imaginemos que deseamos ir más lejos y a partir de imágenes queremos saber si la imagen es de un pez cirujano amarillo o es cualquier otro animal acuático. Entonces ya no basta con saber diferenciar si es un pez o no, o si es amarillo o no, ahora es necesario tener un mayor conjunto de características para saber reconocerlos de manera adecuada en las imágenes, tomando en cuenta tal vez el tamaño, la forma, etc.

A pesar de que los humanos realizamos la tarea de clasificar individuos automáticamente, implementarlo de forma computacional no es una tarea sencilla, justamente porque las características que elegimos influyen directamente en el comportamiento de nuestro clasificador. Además no es fácil decidir cuáles son las características más importantes para que nuestra clasificación tenga los resultados deseados.

2.1.1 Clasificación supervisada y no supervisada

En general, existen dos tipos de clasificación, la supervisada y la no supervisada. En la primera se utiliza un *conjunto de entrenamiento* para poder establecer cómo se separarán las clases, es decir, a partir de un conjunto de individuos ya clasificados se genera el modelo con el que están clasificados, además se espera que el modelo generado así sea capaz de asignar clases correctamente a nuevos individuos que no se encontraban en el conjunto de entrenamiento. La elección adecuada de las características es un paso crucial para la generación del modelo apropiado.

Regresando al ejemplo anterior, si se cuenta con 1000 imágenes de animales acuáticos, se podrían utilizar 200 de ellas ya clasificadas como conjunto de entrenamiento mediante el cual se podría

establecer el modelo que permite separar las 2 clases deseadas (peces cirujano amarillo y el resto de animales acuáticos), y posteriormente aplicar este modelo a las 800 imágenes restantes.

En la clasificación no supervisada ya no se cuenta con el conjunto de entrenamiento, en realidad se espera que el algoritmo sea capaz de encontrar las clases adecuadas por sí mismo. Cuando en la clasificación no supervisada se busca que las clases encontradas maximicen la similitud entre objetos de la misma clase y minimicen la similitud entre objetos de diferentes clases se conoce como agrupamiento (clustering), y cada clase es una agrupación (cluster). Considerando el ejemplo previo, en el caso de clasificación no supervisada se esperaría que el algoritmo sea capaz de separar en diferentes clases las 1000 imágenes, y que una de esas clases sea la de los peces cirujano amarillo.

2.1.2 Espacio de características

En el presente trabajo se usará un clasificador de tipo supervisado, ya que es más sencillo de implementar que un no supervisado. Para la clasificación supervisada se mencionó que cada individuo tiene un conjunto de características que lo describen y el clasificador asignará una clase a individuos nuevos dependiendo del modelo obtenido previamente en el conjunto de entrenamiento. Usualmente es posible cuantificar las características de un individuo, con esto se tiene un vector de características que se encuentra en un espacio de características [37], claramente dependiendo del número de estas será el tamaño del espacio sin embargo para nosotros solo es posible visualizar máximo tres características al mismo tiempo, ver Figura (3.1).

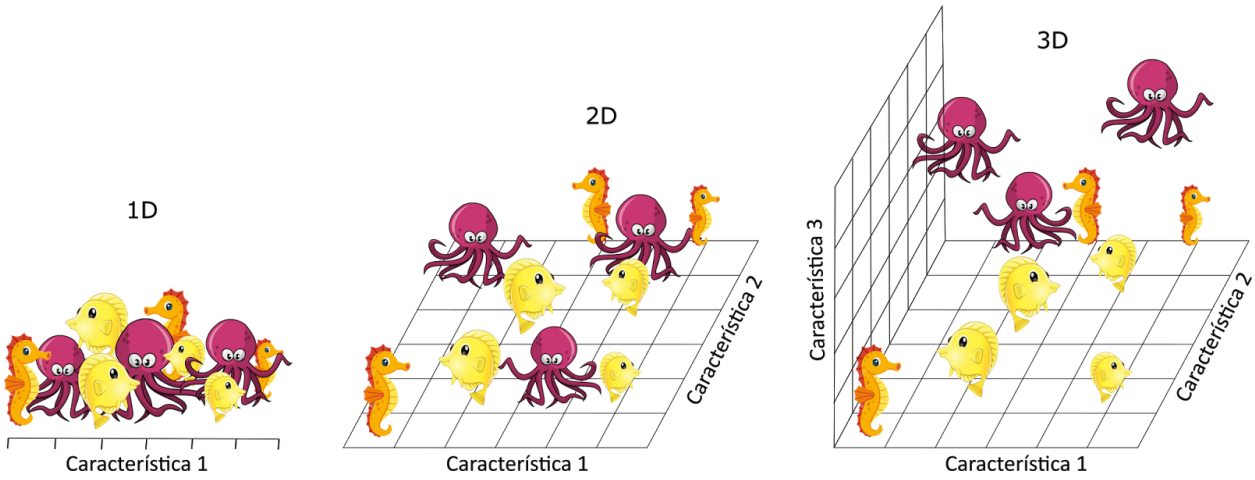

Figura 2.1 Se muestra el ejemplo de los animales acuáticos y sus espacios de características en 1D, 2D y 3D. Cada animal acuático ocupa un lugar específico en el espacio de características, se espera que en algún espacio sea posible separar mediante alguna(s) superficie(s) los diferentes tipos de animales. Imagen inspirada en [38]

En este espacio de características es donde usualmente se implementa el clasificador, en el caso de 1D dividimos el espacio en clases eligiendo puntos, en el caso de 2D se pueden usar rectas o curvas, en el espacio de 3D se pueden usar planos o superficies y en más dimensiones hiperplanos.

Las características que se consideran para implementar un clasificador dependen del tipo de problema que se va a abordar, más adelante se describirán en mayor detalle algunos puntos que se deben tener en cuenta para elegirlas y generar el modelo de clasificación.

2.1.3 Clasificadores generativos y discriminativos

Los clasificadores supervisados se basan fuertemente en la estadística, básicamente tratan de construir un modelo de la distribución de probabilidad de las clases en función de las características [39] con base en el conjunto de entrenamiento, a partir de este modelo se genera el clasificador que asignará una clase a individuos cuyas características conocemos.

El modelo de distribución de clases en función de las características está dado por la función de densidad de probabilidad condicional $f(y|x)$, que básicamente es la probabilidad de ocurrencia de "y" (las clases) bajo la condición "x" (las características). Para producir este modelo de distribución se usan dos tipos de clasificadores, los generativos y los discriminativos. Los primeros generan un modelo de función de densidad de probabilidad conjunta $f(x,y) = f(x|y) \cdot f(y)$ y usando el teorema de Bayes [40] obtienen la distribución de probabilidad condicional $f(y|x)$. Los discriminativos obtienen $f(y|x)$ de forma directa [41]. Ejemplos de clasificadores generativos son el análisis discriminante Gaussiano (GDA por sus siglas en inglés) y el clasificador de Naive Bayes, mientras que la regresión logística y las redes neuronales son ejemplos de clasificadores discriminativos.

En este trabajo se hace uso de la regresión logística como clasificador debido a que como se ha visto es un método más directo para obtener la función $f(y|x)$, por esta razón el apartado siguiente se dedica a su estudio.

2.2 Regresión logística

Antes de abordar el problema de clasificación mediante regresión logística en esta sección se dará una introducción teórica para comprender las bases usando como ejemplo la regresión lineal para después pasar al caso de la regresión logística.

2.2.1 Regresión Lineal

Después de realizar algún experimento es común buscar la forma en que se relacionan las variables independientes (las que controlamos) y las variables dependientes (los valores que se obtienen después de una medición), esto con el fin de encontrar un modelo que describa el comportamiento de nuestro objeto de estudio y poder predecir nuevos resultados al extender nuestro dominio de las variables independientes.

Un método estadístico que permite obtener estas relaciones es el análisis de regresión, el cual en el caso más sencillo permite encontrar la recta que "mejor se ajusta" a un conjunto de datos (X: Variable independiente, Y: variable dependiente), y en general se puede extender a casos en los que se cuenta con más de una variable independiente o bien ajustar otro tipo de curvas .

Es importante considerar que en un modelo de regresión hay una distribución de probabilidad para cada valor de X y los promedios de estas distribuciones de probabilidad cambian con X siguiendo alguna tendencia [42]. Esto se ilustra en la Figura (3.2):

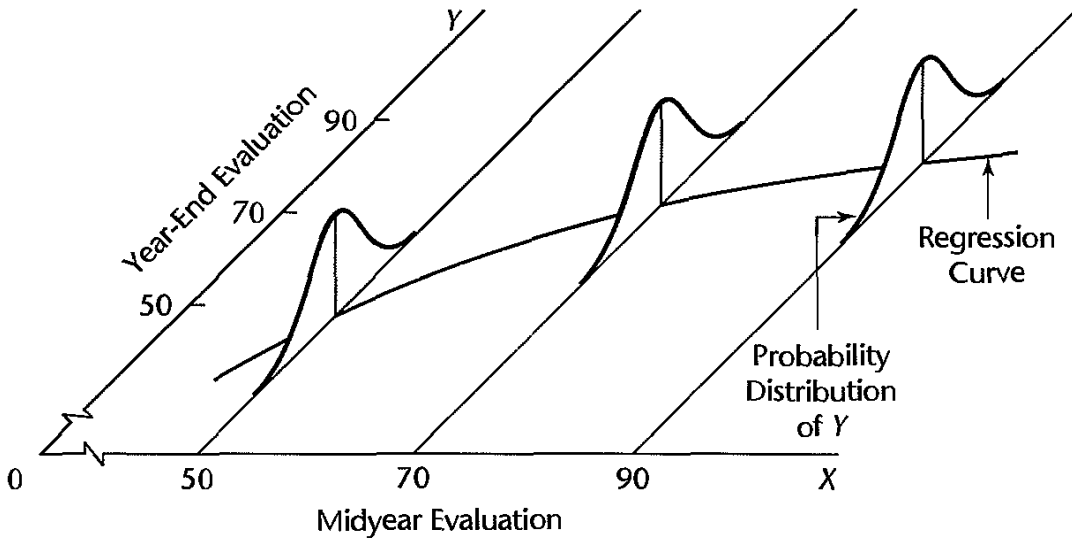

Figura 2.2 Representación gráfica de un modelo de regresión, se muestran las distribuciones de probabilidad para algunos valores de X. Imagen tomada de [42].

Usualmente, para encontrar la recta que mejor se ajusta a un conjunto de pares de datos se utiliza el método de mínimos cuadrados, en el cual se minimiza la función Q generada al restar los cuadrados de las diferencias entre valores obtenidos experimentalmente (Y) y los valores generados mediante la regresión ($\widehat{Y}$), esto nos permite hallar los coeficientes lineal β_1 e independiente β_0 de la recta que estamos buscando:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

En esta ecuación los valores van de i=1 hasta n, donde n es el número de mediciones que se han realizado en el experimento, $\beta_0 - \beta_1 x_i$ es la recta deseada, (x_i, y_i) son los valores que toman las variables (X,Y). Al final se desea que Q sea mínima (i.e. el error sea mínimo), entonces para encontrar los valores de β_0 y β_1 se minimiza la expresión para Q.

Aunque el método de mínimos cuadrados es el más usado para estimar los parámetros β_0 y β_1 de relaciones lineales entre variables, existe uno más general conocido como *método de máxima verosimilitud*, el cual maximiza la *función de verosimilitud*, que es la densidad de probabilidad de los datos observados (Y), pero en función de los parámetros desconocidos (β_0 y β_1) [43]. Sin embargo para hacer uso del método de máxima verosimilitud es necesario conocer la forma de la distribución de los datos observados.

Para el caso de la regresión lineal, el modelo general establecido es el siguiente:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (2.1)$$

Donde:

y_i es el i-ésimo valor que toma la variable dependiente Y

x_i es el i-ésimo valor que toma la variable independiente X

ε_i es el i-ésimo término de error aleatorio

β_0 y β_1 son los parámetros a buscar

Para conocer la forma de la distribución de los datos se asume que los términos aleatorios ε_i poseen una distribución normal con promedio $E[\varepsilon_i] = 0$ y varianza constante $Var[\varepsilon_i] = \sigma^2$, denotando su distribución como $\varepsilon_i \sim N(0, \sigma^2)$.

Al hacer estas consideraciones, y debido a:

a) Las propiedades del valor esperado de una variable aleatoria [44], y

b) Que el término $\beta_0 + \beta_1 x_i$ es un valor constante

La ecuación (2.1) permite obtener el siguiente resultado:

$$E[y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = E[\beta_0 + \beta_1 x_i] + E[\varepsilon_i] = \beta_0 + \beta_1 x_i$$

Considerando que la función de regresión se obtiene a través de los promedios de las distribuciones de probabilidad, el resultado anterior se puede generalizar para obtener la función de regresión del modelo descrito en la ecuación (2.1):

$$E[Y] = \beta_0 + \beta_1 X \quad (2.2)$$

Además, tomando en cuenta los incisos anteriores a) y b) se puede concluir de manera similar que cada distribución de probabilidad de Y (es decir cada y_i) es normal, con media dada por la ecuación (2.2) y varianza constante igual a la del error, entonces cada y_i posee una distribución denotada por $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

De esta manera, al conocer la distribución del error se encontró la distribución de las observaciones (la variable dependiente Y), con esto es posible encontrar la función de verosimilitud, ya que la densidad de probabilidad para una sola observación y_i está dada por [42] (considerando la distribución de probabilidad normal encontrada antes):

$$f_i(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right]$$

En la ecuación anterior $f_i(y_i|x_i;\beta)$ expresa la distribución de probabilidad de y_i condicionada por x_i, parametrizada por β (que denota a β_0 y β_1).

Entonces para n observaciones se tendrá que la función de verosimilitud L es el producto de las n funciones f_i (debido a la independencia):

$$L(\beta_0, \beta_1, \sigma^2) = f(Y|X;\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2]$$

Que puede reducirse a:

$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2] \quad (2.3)$$

Esta ecuación es la que se debe maximizar para encontrar los valores estimados $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{\sigma}^2$ (la varianza también se considera ya que suele ser desconocida).

Cuando se maximiza la ecuación (2.3) los valores estimados para $\widehat{\beta}_0$, $\widehat{\beta}_1$ dan el mismo resultado que al usar el método de mínimos cuadrados. Además es importante destacar que al usar el método de mínimos cuadrados se está asumiendo que el error tiene una distribución $N(0, \sigma^2)$ [45], algo que no siempre se cumple como se verá en el caso de la regresión logística.

2.2.2 Regresión Logística

Después de haber revisado el caso de la regresión lineal podemos abordar de mejor forma el caso en el que nuestra observación sólo puede tomar valores discretos, por ejemplo, si en un experimento se desea determinar si una persona pertenece al género masculino o femenino dependiendo de su estatura, la variable independiente es la estatura mientras que la variable dependiente es el género, que puede ser codificado con 1 si es mujer y 0 si es hombre. Aunado a esto puede haber casos donde se cuente con más de una variable independiente, por ejemplo en cuestiones médicas, si se desea saber si una persona está o no enferma considerando datos como la edad, presión arterial, nivel de glucosa, etc.

En este tipo de problemas buscar una solución de tipo lineal no parece tener sentido ya que si para una primera inspección se grafican los datos se tienen representaciones del estilo de la Figura (2.3), que corresponde al ejemplo antes descrito:

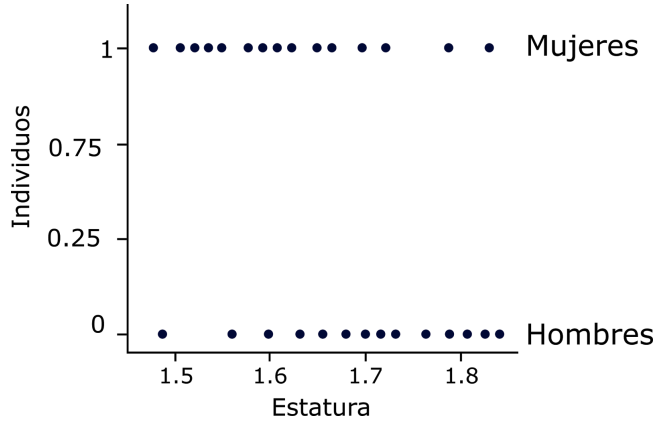

Figura 2.3 Gráfica de puntos para un experimento hipotético donde se tiene el género y estatura de una persona, y se ha codificado con 1 a las mujeres y con 0 a los hombres, se observa que la gráfica no proporciona suficiente información para poder hacer predicciones del género a partir de la estatura.

De la Figura (2.3) sólo se puede inferir que a mayor estatura es más probable que la persona sea del género masculino y viceversa, sin embargo no es posible contar con una recta que nos permita hacer predicciones como en el caso de la regresión lineal. Para conocer mejor la tendencia de este tipo de datos es posible obtener un valor aproximado del promedio $E[Y]$ como en el caso de la regresión lineal, y esto se logra creando intervalos dentro de la variable dependiente y calculando el promedio de la variable de salida dentro de cada intervalo [46]. Haciendo esto se obtienen curvas similares a la Figura (2.4):

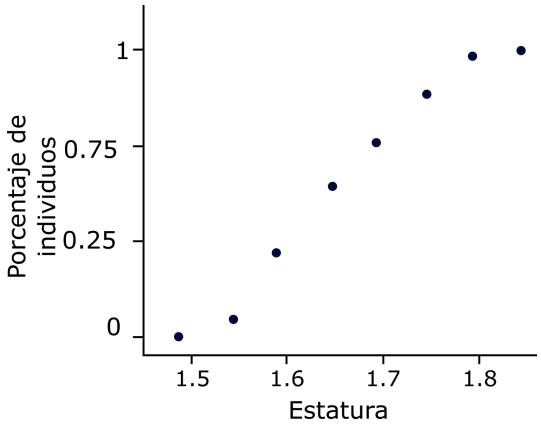

Figura 2.4 Gráfica del porcentaje de individuos que caen dentro de un rango de estaturas para el experimento hipotético de la Figura (2.3).

De la Figura (2.4) se puede observar que emerge una tendencia mucho más clara que la observada en la Figura (2.3). Es importante advertir que en este tipo de gráficas los valores aproximados de $E[Y]$ nunca serán menores a 0 ni mayores 1, esto ocurre debido a la forma en que hemos calculado los promedios $E[y_i]$, además en estas gráficas se tendrán asíntotas en esos valores, haciendo que en la porción central de la variable independiente (estatura en el ejemplo) se tenga un comportamiento casi lineal, mientras que en los extremos se tenga un comportamiento curvo.

La forma particular de la curva así generada (de S) es muy similar a las encontradas en las funciones de distribución de una variable aleatoria, ver Figura (2.5).

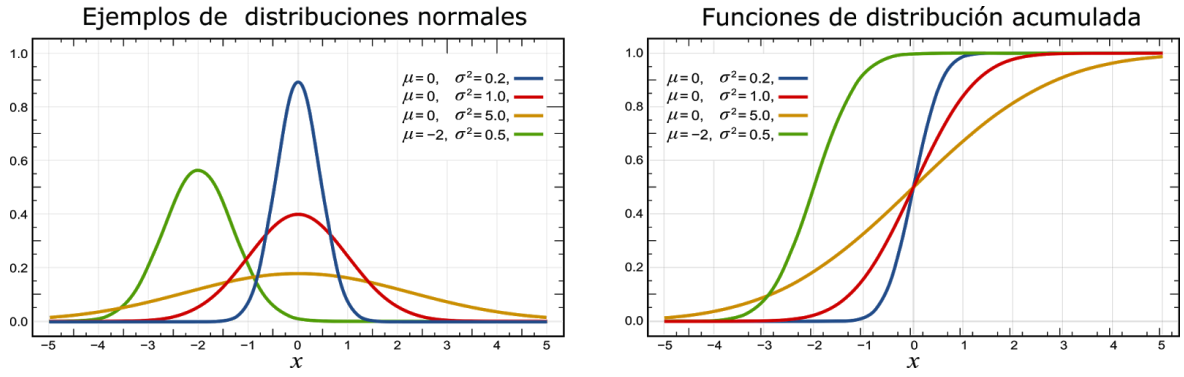

Figura 2.5 Ejemplos de distribuciones normales y sus correspondientes funciones de distribución acumulada, nótese la forma de "S" de éstas últimas. Tomada de [47].

Una de las distribuciones más usadas para modelar $E[Y]$ en este tipo de problemas es la distribución logística, cuya función está definida por:

$$f(z) = \frac{1}{1+e^{-z}} \tag{2.4}$$

Y gráficamente se puede observar en la Figura (2.6):

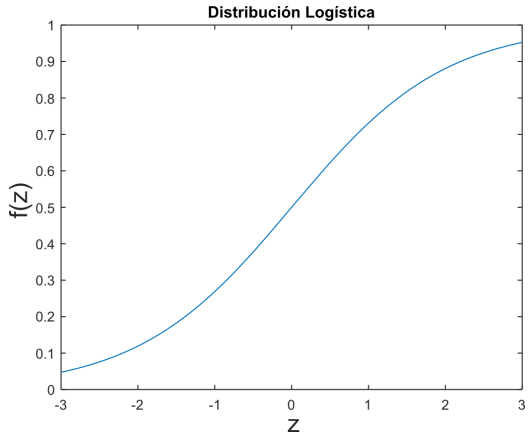

Figura 2.6 Gráfica de la distribución logística, ecuación (2.4).

Las razones por las que se suele usar esta función es porque está definida entre 0 y 1, lo cual no siempre es posible para otros modelos, es bastante sencilla y flexible, y es muy utilizada en el ámbito clínico ya que se considera que z es un índice que usa diferentes contribuciones de diferentes factores de riesgo de una enfermedad y $f(z)$ representa el riesgo para un valor dado de z [48].

En el modelo logístico z es una combinación lineal de los parámetros desconocidos: $z = \beta_0 + \beta_1 X$, de esta forma sustituyendo en la ecuación anterior resulta:

$$f_\beta(x_i) = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} \tag{2.5}$$

Ahora se tiene un problema similar al de la regresión lineal donde deseamos ajustar la función $f_\beta(x_i)$ basándonos en nuestros datos X y la salida dicotómica 0 ó 1, y desconocemos los parámetros β_0 y β_1. Tomando como base la ecuación (2.1), podríamos usar un modelo similar para el caso de la regresión logística, teniendo:

$$y_i = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} + \varepsilon_i \tag{2.6}$$

Se puede aplicar el operador de valor esperado en ambos lados de la ecuación anterior, asumiendo que se cumple que $E[\varepsilon_i] = 0$, de forma similar a como se hizo en la regresión lineal, y como el primer término del segundo miembro es una constante, el resultado de aplicar el operador es:

$$E[y_i] = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}$$

Por otra parte, si la probabilidad de que $y_i = 1$ es π_i, entonces la probabilidad de que $y_i = 0$ es $1 - \pi_i$, y considerando que el valor esperado de una variable aleatoria está dado por la suma de los valores que toma la variable aleatoria multiplicados por su probabilidad, se tiene que:

$$E[y_i] = (1) * (\pi_i) + (0) * (1 - \pi_i) = \pi_i$$

De estas dos últimas ecuaciones se puede concluir lo siguiente:

$$E[y_i] = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} = \pi_i \tag{2.7}$$

Es decir, el valor esperado de la variable aleatoria Y es igual a la probabilidad de que esa variable aleatoria sea igual a 1.

Con estos resultados se analizará la distribución del error, para mostrar porqué no es posible usar en la regresión logística el método de mínimos cuadrados.

Como la variable dependiente y_i puede tomar solo 2 valores, 0 o 1, de la ecuación (2.6) el error puede tomar los valores $\varepsilon_i = -E[y_i]$ ó $\varepsilon_i = 1 - E[y_i]$ para y=0 ó 1 respectivamente, por lo que ya no se cumple que la distribución del error sea normal como en el caso de la regresión lineal.

Por otro lado se obtendrá el valor de la varianza del error, despejando ε_i de la ecuación (2.6), usando la ecuación (2.7) y aplicando el operador de varianza de ambos lados se obtiene que: $\sigma^2[\varepsilon_i] = \sigma^2[y_i - \pi_i] = \sigma^2[y_i]$ ya que π_i es un valor constante.

Por el resultado anterior se requiere encontrar el valor de la varianza de y_i, considerando que la varianza se puede obtener mediante: $\sigma^2[Z] = E[Z] - (E[Z])^2$ y de la ecuación (2.6), resulta:

$$\sigma^2[y_i] = E[\frac{1}{1+e^{-\beta_0-\beta_1 x_i}} + \varepsilon_i] - (E[\frac{1}{1+e^{-\beta_0-\beta_1 x_i}} + \varepsilon_i])^2 = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} - (\frac{1}{1+e^{-\beta_0-\beta_1 x_i}})^2$$

De esto se deduce que la varianza del error depende de x_i, dejando de ser constante como se había supuesto para el caso de la regresión lineal.

Con ambos resultados se observa que el error dejó de ser una variable aleatoria con distribución normal y varianza constante, por lo que el uso del método de mínimos cuadrados en esta situación ya no es aplicable.

Como se mencionó en el caso de la regresión lineal, si se desea aplicar el método de máxima verosimilitud es necesario conocer la densidad de probabilidad de las observaciones, sin embargo a partir de lo que se ha observado hasta ahora es fácil concluir que la distribución de Y es binomial, ya que cumple con las siguientes características [44]:

- Y es una variable aleatoria que toma dos posibles valores, usualmente codificados como 1 (éxito) ó 0 (fracaso).
- En cada ensayo (medición) hay una probabilidad P de éxito y una probabilidad $1 - P$ de fracaso.
- La probabilidad P de éxito permanece constante en cada ensayo.
- Los resultados de ensayos sucesivas son estadísticamente independientes entre sí.

Conociendo el tipo de distribución de las observaciones es posible obtener la función de verosimilitud, primero se requiere obtener la distribución de una observación y_i, como se tiene una distribución binomial entonces $f_i(y_i|x_i;\beta) = \pi_i^{y_i}[1-\pi_i]^{1-y_i}$ [43], si se tienen n observaciones y considerando que son independientes la función de verosimilitud L se obtiene mediante la multiplicación de las n distribuciones f_i:

$$L(\beta_0,\beta_1) = f(Y|X;\beta) = \prod_{i=1}^{n} \pi_i^{y_i}[1-\pi_i]^{1-y_i} \tag{2.8}$$

Esta función es la que se debe maximizar para encontrar los valores de los parámetros β, en el apéndice 1 se muestra cómo se encuentra esta función L de forma explícita para β, resultando en:

$$ln[L(\beta)] = \sum_{i=1}^{n}[y_i \cdot \{\beta_0 + \beta_1 x_i\} - ln\{1 + e^{\beta_0+\beta_1 x_i}\}] \tag{2.9}$$

La función logaritmo se aplicó porque es más sencillo trabajar con esta expresión cuando se maximiza, además maximizar la ecuación (2.8) da el mismo resultado que maximizar la ecuación (2.9). De esta forma maximizando la ecuación (2.9) se obtienen las siguientes expresiones para los valores de β (ver apéndice 2):

$$\sum_{i=1}^{n}[y_i - \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}] = 0 \tag{3.10}$$

$$\sum_{i=1}^{n} x_i \cdot [y_i - \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}] = 0 \tag{3.11}$$

A diferencia de la regresión lineal, las ecuaciones así obtenidas son no lineales en los parámetros β que estamos buscando, por esta razón usualmente se utilizan algoritmos numéricos iterativos para resolverlas.

2.3 Clasificación Mediante Regresión Logística

Hasta ahora la regresión logística nos permite predecir la probabilidad de que una variable aleatoria y_i tome el valor codificado como 1 (éxito) ya que solo hemos realizado el ajuste de nuestros datos con la función de distribución logística, sin embargo para el caso de la clasificación se requiere que a la salida se cuente con la clase (1 ó 0) y no con la probabilidad de éxito o fracaso, así que es necesario transformar esta salida a una clase.

Consideremos de nuevo la ecuación (2.7), $\pi_i = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}$, indica que la probabilidad de que $y_i = 1$ está dada por la distribución logística (2.5) que varía de 0 a 1, considerando que es igual de probable que nuestra nuestra variable aleatoria $Y = y_i$ tome el valor 0 ó 1 podemos esperar lo siguiente, de la ecuación (2.5):

$$\text{Si } f_\beta(x_i) \geq 0.5 \;\Rightarrow\; y_i = 1 \text{, i.e. } y_i \text{ es de clase 1}$$

$$\text{Si } f_\beta(x_i) < 0.5 \;\Rightarrow\; y_i = 0 \text{, i.e. } y_i \text{ es de clase 0}$$

De la figura (2.6) podemos observar que $f_\beta(x_i) \geq 0.5 \Leftrightarrow \beta_0 + \beta_1 x_i \geq 0$, aplicando el mismo razonamiento a la segunda condición anterior se obtiene lo siguiente:

$$\text{Si } \beta_0 + \beta_1 x_i \geq 0 \;\Rightarrow\; y_i = 1 \text{, i.e. } y_i \text{ es de clase 1}$$

$$\text{Si } \beta_0 + \beta_1 x_i < 0 \;\Rightarrow\; y_i = 0 \text{, i.e. } y_i \text{ es de clase 0}$$

Con esto se ha encontrado la recta $B = \beta_0 + \beta_1 X$ que dividirá las dos clases en nuestro espacio de características, conocida como frontera de decisión, y como se mencionó antes dependiendo de la dimensión del espacio de características se tendrá la divisón de las clases mediante un punto, una recta, un plano o un hiperplano. Esto se muestra en la Figura (2.7):

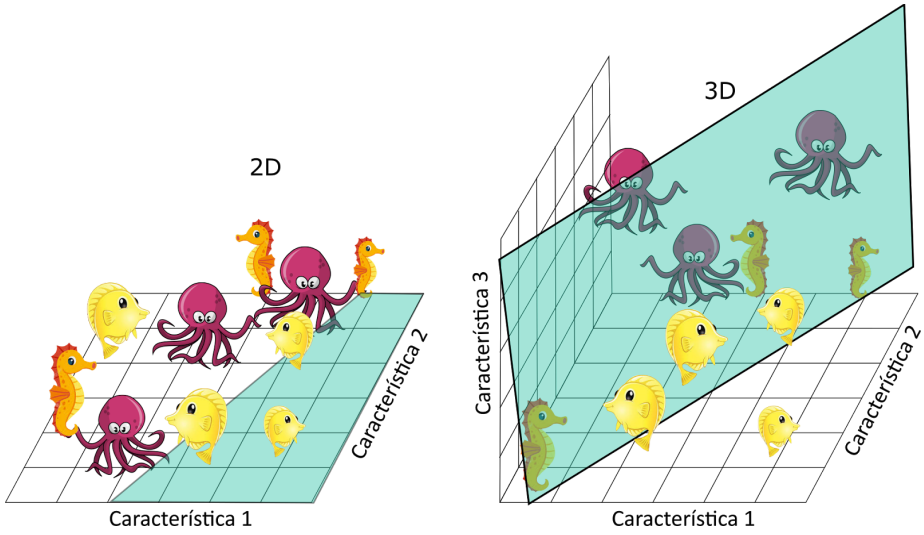

Figura 2.7 Ejemplos de frontera de decisión para dos espacios de características en 2D y 3D.

Es importante aclarar que aunque hasta ahora se han asumido combinaciones lineales en la característica X: $B = \beta_0 + \beta_1 X$, esto no es un requisito indispensable ya que se pueden tener otras formas, ver figura (2.8), por ejemplo polinomios, parábolas, círculos: $B = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2$, en este ejemplo se tienen dos características X_1 y X_2 por lo que el espacio de características es de 2D. Al final la forma que adoptará la frontera de decisión dependerá de cómo está distribuido el conjunto de

entrenamiento en el espacio de características, y claramente es mucho más sencillo trabajar máximo hasta con 3 características para poder visualizar la frontera de decisión, sin embargo existen muchos problemas en los que se cuenta con un número enorme de características haciendo imposible la visualización de la frontera, en estos casos se requieren otras formas de analizar si la forma que estamos adoptando para la frontera es la más adecuada o no.

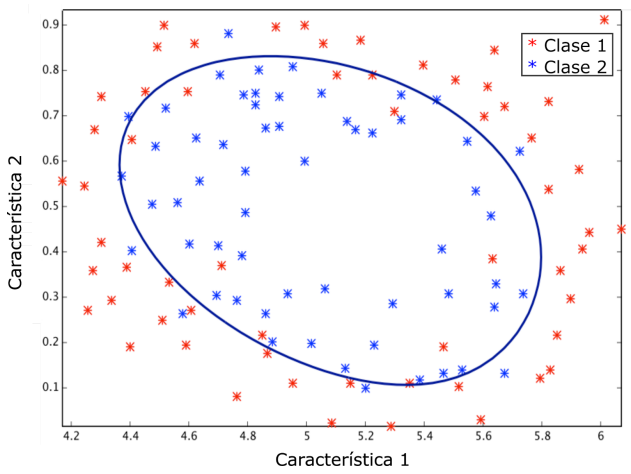

Figura 2.8 Ejemplo de frontera de decisión para un espacio de características en 2D que tiene una forma no lineal. Imagen tomada de [49].

2.3.1 Reducción y normalización de las características

Elegir de manera adecuada las características que describirán a nuestros individuos es una tarea crucial en el proceso de clasificación, incluso puede ser más importante que el mismo algoritmo de clasificación [50], por esta razón es importante tener conocimiento del tipo de problema de clasificación que se desea abordar y así determinar las características que permitirán realizar buenas predicciones de las clases.

Algunos de los puntos que hay que tomar en cuenta al momento de elegir las características se listan a continuación [37]

- Hay que buscar características que permitan discriminar de la mejor forma las clases de los objetos estudiados, es decir hay que tratar de conocer a fondo el objeto de estudio, y además que no sean redundantes, ya que se traduce en uso innecesario de recursos de tiempo y almacenamiento.
- Hay que buscar el número mínimo necesario de características para que el clasificador funcione adecuadamente, esto se detalla más adelante.
- Considerar que el cálculo de las características sea rápido [51], ya que esto se verá reflejado en el rendimiento del clasificador.
- Tener especial cuidado con los datos atípicos, outliers, ya que suelen ocasionar superposición en el espacio de características.

Es importante destacar que tener un número grande de características no implica que el algoritmo de clasificación funcionará de forma adecuada, de hecho existe el llamado efecto Hughes que indica cómo

al aumentar la dimensión del espacio de características (cientos y miles de dimensiones) y al mismo tiempo mantener un tamaño fijo del conjunto de entrenamiento se producirá una gran dispersión de los datos en el espacio de características, esto hará más sencillo encontrar un hiperplano que divida las clases sin embargo producirá sobreajuste, descrito en el siguiente apartado, y ocasionará demasiados errores cuando se busque clasificar nuevos elementos [38]. Por esta razón si se aumenta el número de características es necesario incrementar el número de datos del conjunto de entrenamiento, lo cual no siempre es posible. El efecto del error de estimación debido a la dimensión del espacio de características se muestra en la Figura (2.9), se puede observar que se desea encontrar el mínimo de esa gráfica, que corresponde al error de clasificación mínimo para un número determinado de características [37].

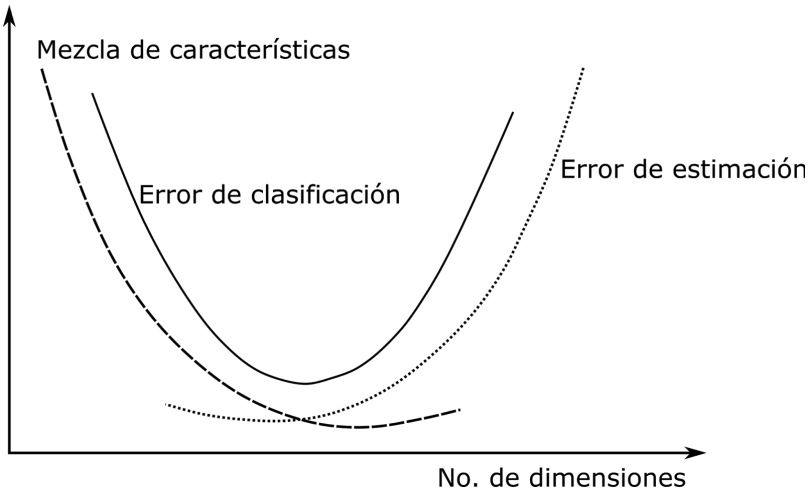

Figura 2.9 El error de clasificación presenta una forma de U en función del número de dimensiones del espacio de características si el tamaño del conjunto de entrenamiento permanece constante. Tomada de [37].

Además de evitar el efecto Hughes, el tener un número razonable de características permite disminuir la complejidad de cómputo, ya que como se vio antes para clasificar nuevos elementos será necesario obtener sus características, y esto ayudará a mejorar tanto el rendimiento del clasificador como del algoritmo encargado de encontrar la frontera de decisión, además esto contribuye a evitar un exceso de espacio de almacenamiento [52].

Ya que el número de características tiene efectos importantes en los clasificadores, se han desarrollado algoritmos que permiten reducir la dimensión de este espacio, y se dividen en dos tipos [53]:

- Selección de características, consiste en seleccionar un subconjunto de las características de entrada, eliminando la redundancia entre las mismas.
- Extracción de características, son métodos que crean nuevas características usando transformaciones o combinaciones del conjunto de características original, las nuevas características suele tener una interpretación más complicada.

Para finalizar esta sección se discutirá brevemente la normalización de las características. Ya que usualmente las características poseen diferentes unidades de medida, tienden a encontrarse en diferentes rangos dentro del espacio de características, sin embargo esto puede no ser del todo deseable

para algunos algoritmos numéricos encargados de encontrar los parámetros β (ver la sección dedicada a la regresión logística), por esta razón muchas veces se busca que las características se encuentren dentro de rangos similares y para esto se hace uso de la normalización, con la cual se remueven las unidades de las mismas, la forma más sencilla de realizar esto es mediante la siguiente transformación:

$$\widehat{W}_i = \frac{w_i - \mu_w}{\sigma_w}$$

Donde w_i es el i-ésimo valor que toma la característica W, μ_w es el promedio de la característica W, σ_w es la desviación estándar de la característica W y $\widehat{w}_i$ es el i-ésimo valor normalizado que toma la característica W. Esta transformación produce características con promedio cero y desviación estándar uno.

Sin embargo se debe tener cuidado al utilizar esta normalización, ya que aunque en algunos problemas permite una clara separación entre clases dentro del espacio de características, a veces tiene el efecto contrario [37].

2.3.2 Elección del modelo y validación del clasificador

Debido a que usualmente es necesario probar diferentes modelos de fronteras de decisión, sobre todo cuando se tiene un gran número de características y no es posible visualizarlas fácilmente en una gráfica, lo que se suele hacer es dividir el conjunto de objetos clasificados en tres partes , se recomienda que el 60% de los datos se utilicen para el conjunto de entrenamiento, 20% para probar con diferentes modelos de frontera (usando diferentes grados de polinomios sobre las características), llamado conjunto de validación cruzada, y 20% para estimar el error del clasificador después de haber elegido el tipo de frontera, llamado conjunto de prueba. Antes de dividir el conjunto de datos principal es necesario mezclarlos de manera aleatoria para que no exista una tendencia en el desarrollo del clasificador que pueda interferir en los resultados.

De esta forma el procedimiento para generar un clasificador es el siguiente:

1. Crear un conjunto de diferentes modelos de frontera de decisión, por ejemplo añadiendo términos no lineales.
2. Encontrar los valores de los parámetros β para cada modelo generado en el paso 1 usando el conjunto de entrenamiento.
3. Hallar el modelo con el error mínimo usando el conjunto de validación cruzada.
4. Estimar el error del clasificador usando el conjunto de prueba, usando el modelo hallado en el paso 3.

El error entre un valor estimado $\hat{y}_i$ y su valor verdadero y_i se puede calcular mediante la siguiente expresión:

$$err(\hat{y}_i, y) = \begin{cases} 1 \text{ si } \hat{y}_i = 1 \text{ y } y = 0 \text{ ó } \hat{y}_i = 0 \text{ y } y = 1 \\ 0 \text{ en otro caso} \end{cases}$$

Con esto, el error sobre uno de los conjuntos se puede evaluar mediante:

$$error = \frac{1}{n}\sum_{i=1}^{n} err(\hat{y}_i, y)$$

Donde n es el tamaño del conjunto utilizado.

Concluyendo con esta parte, existen dos problemas bajo los cuales un clasificador puede tener un muy mal desempeño llamados subajuste y sobreajuste. Gráficamente ambos se muestran a continuación:

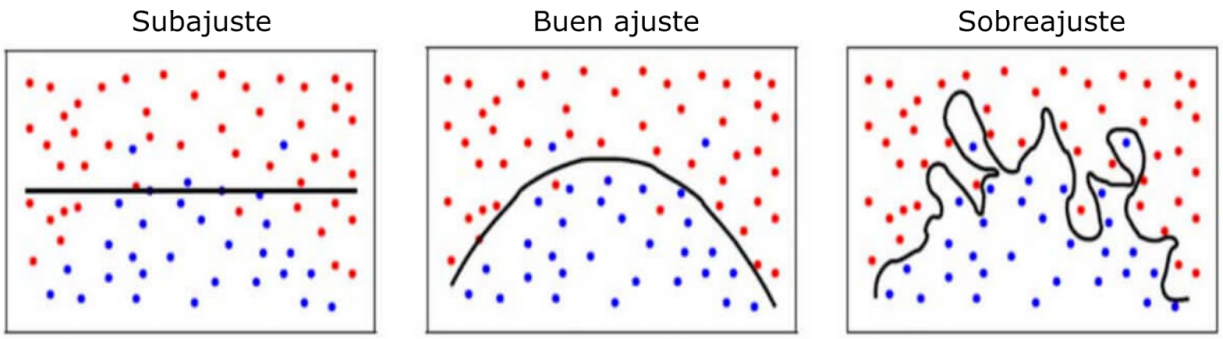

Figura 2.10 Se muestra un mismo espacio de características con tres fronteras de decisión diferentes, una presenta subajuste, oyta un ajuste adecuado y la otra sobreajuste, la primera y la última provocarán problemas en el clasificador. Tomada de [54].

Como se puede observar de la Figura (2.10) en el caso del subajuste el modelo de la frontera de decisión describe de manera muy deficiente el comportamiento de los datos clasificados, esto es ocasionado por el uso de un modelo demasiado simple (en este caso lineal) o el uso de muy pocas características, mientras que en el caso del sobreajuste el modelo de la frontera de decisión se ajusta demasiado bien a los datos en el conjunto de entrenamiento, sin embargo al tratar de generalizar este modelo a datos no clasificados se tendrá un rendimiento muy malo, es ocasionado por contar con demasiadas características o un modelo demasiado complejo.

Debido a las causas por las que se presentan ambos problemas, algunas soluciones a ellos se muestran a continuación:
Para corregir el subajuste:

- Conseguir un mayor número de características.
- Hacer más complejo el modelo de la frontera de decisión.

Para corregir el sobreajuste:

- Conseguir un mayor número de ejemplos de entrenamiento.
- Usar un menor número de características.

Sin embargo cuando se tienen más de tres características y se tiene un mal desempeño del clasificador, es difícil saber si se debe a un subajuste o sobreajuste del modelo, por lo que existen las llamadas curvas de aprendizaje que nos permiten diferenciar cuál de los dos problemas está presente en el clasificador diseñado:

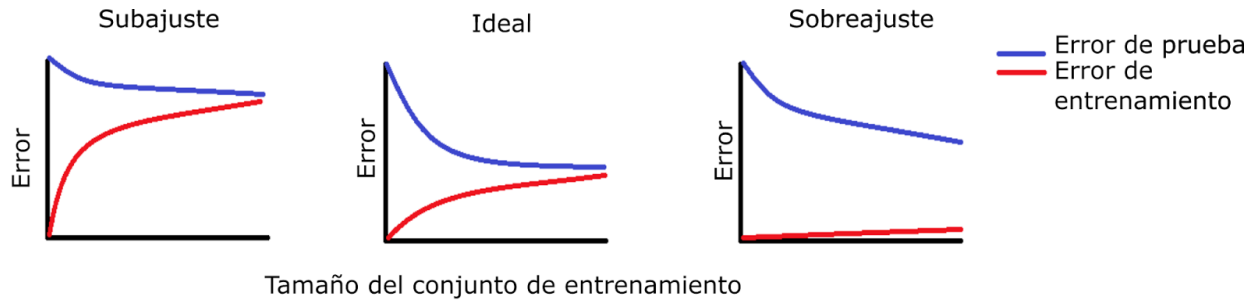

Figura 2.11 Se muestra ejemplos de curvas de aprendizaje generadas ´para clasificadores que presentan subajuste, sobreajuste y el caso ideal. Tomada de [55].

En estas gráficas se tiene el comportamiento del error vs el tamaño del conjunto de entrenamiento (m) para un modelo fijo de frontera de decisión, el gráfico de la izquierda es el de subajuste, ya que al inicio se tienen pocos datos en el conjunto de entrenamiento el error es prácticamente cero (curva roja) y conforme se van añadiendo datos al conjunto de entrenamiento el error va aumentando, sin embargo el error para el conjunto de prueba va disminuyendo conforme m aumenta pero se queda de forma estacionaria en un valor de error demasiado grande, esto ocurre porque llegará un punto en el cual aunque se aumenten más ejemplos de entrenamiento el modelo ya no mejorará por su simplicidad.

Por otro lado cuando se presenta un problema de sobreajuste el comportamiento de ambos errores es similar, pero ahora está presente una brecha entre los dos gráficos (ver gráfico derecho de la Figura (2.11)), haciendo que el error del conjunto de entrenamiento sea bajo, pero el error del conjunto de prueba es demasiado alto, esto se debe a lo mencionado anteriormente ya que al tener sobreajuste en el conjunto de entrenamiento el error será pequeño pero al generalizar el error se disparará en sujetos que no estuvieron en el conjunto de entrenamiento.

Al obtener estas curvas de forma experimental el comportamiento no será tal cual el mostrado en la Figura (2.11), sin embargo la tendencia es lo que se estará buscando para determinar si el clasificador implementado sufre de subajuste o sobreajuste.

3.- Detección de células hipofisarias en imágenes de fluorescencia

“There are more cells in your body than there are galaxies in the known universe”
Nicholas Bakalar

En este capítulo se describe la manera en que se realizó la detección de células en los stacks de imágenes adquiridos mediante microscopía de fluorescencia, primero se describirá el tipo de imágenes con las que se trabajó, después se detallará el pre-procesamiento aplicado a ellas, posteriormente se explicará cómo se aplicó un clasificador de regresión logística para la localización de las células y en el siguiente apartado se detallarán los resultados obtenidos.

3.1 Descripción de los stacks de Imágenes

Como se ha mencionado antes los stacks de imágenes con los que se trabajó en este proyecto son obtenidos mediante microscopía de fluorescencia, el tejido a analizar fue el de hipófisis de ratones donde específicamente se deseaba identificar las células contenidas en él. La Figura (3.1) muestra un ejemplo de este tipo de stacks, ya que se toman imágenes a lo largo del tiempo el problema de fotoblanqueamiento visto en el capítulo 2 cobra gran relevancia y en muchas ocasiones es evidente a simple vista al recorrer el stack en el tiempo.

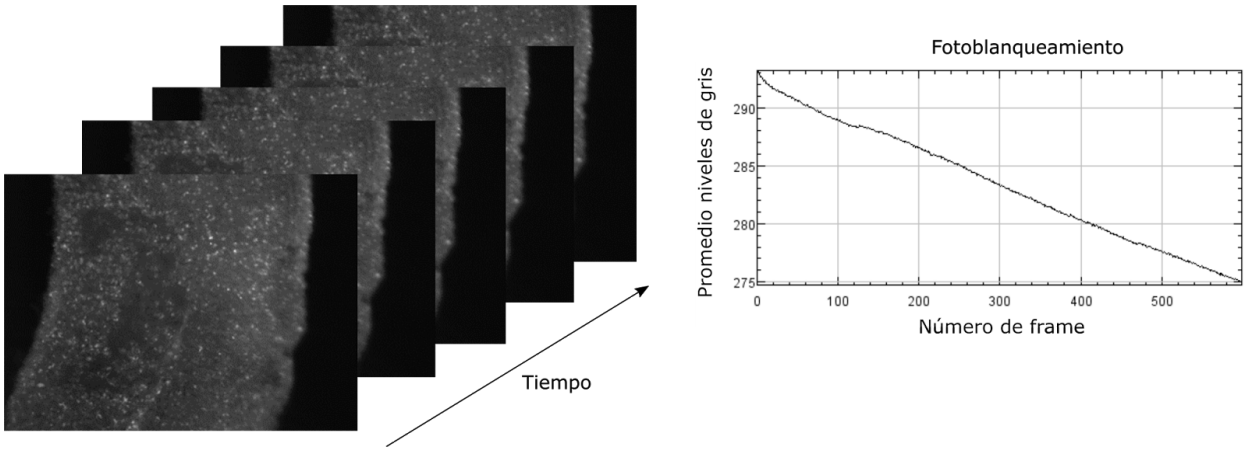

Figura 3.1 Se muestra un ejemplo de stack de imágenes junto con una gráfica del promedio de los niveles de gris de cada frame, es decir el fotoblanqueamiento de las imágenes que integran el stack.

Se trabajó con 12 stacks de imágenes los cuales tienen un ancho que está en el intervalo de los 1392 a los 263 pixeles, un alto que está en el intervalo de los 207-1040 pixeles y un número de frames entre los 300 y los 1350, además el rango dinámico de las imágenes más pequeñas es de 8 bits y el resto es de 16 bits.

Para localizar las células en los stacks el experto utiliza el software *ImageJ* o *Fiji [56]*, el segundo solo es una extensión del primero, que permiten el análisis y procesamiento de imágenes y son de distribución libre. Como se puede observar en la Figura (3.2) la identificación de las células no se basa en los conjuntos de pixeles brillantes que se observan fácilmente en la figura, y que suelen corresponder a células muertas, en realidad el experto va recorriendo el stack en el tiempo y busca posiciones en las que exista un conjunto de pixeles con un aumento gradual en la intensidad y una posterior disminución, aunque muchas veces dichos cambios se presentan más de una vez y de forma muy rápida y es claro que las variaciones no son debidas al fotoblanqueamiento. El experto se suele ayudar de la aplicación de pseudocolor sobre los stacks para distinguir de mejor forma dichos cambios ya que muchas veces no es fácil distinguir las células con stacks en escala de grises.

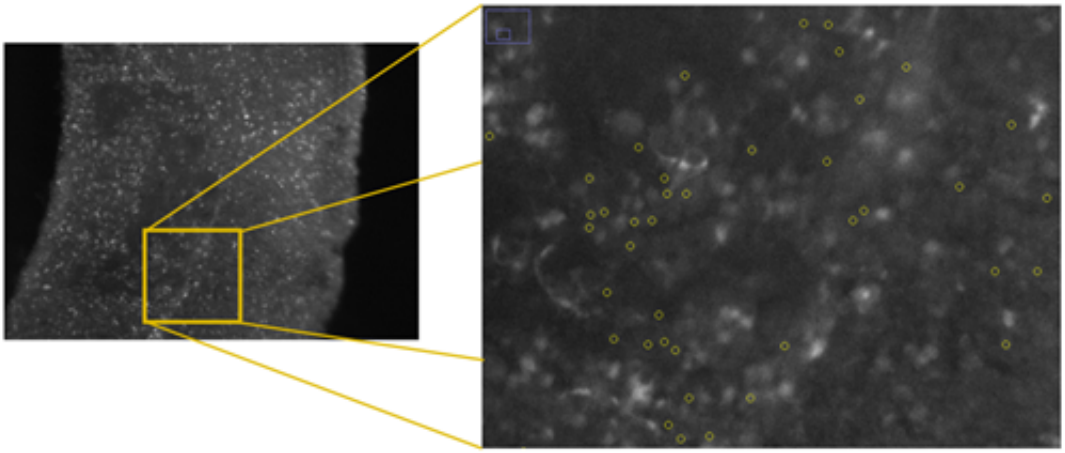

Figura 3.2 Se muestra un ejemplo de detección de células por el experto en un stack de imágenes, los círculos indican zonas donde hay un aumento en la intensidad con una posterior disminución en el eje temporal del stack, los puntos brillantes suelen indicar células muertas.

Cuando el experto ha distinguido algún conjunto de pixeles que corresponda a una célula encierra dicha zona dentro de un círculo, el radio del círculo va a depender del tamaño esperado de las células ya que la escala en las imágenes dependerá de la forma en que se hayan adquirido, cabe resaltar que en los stacks analizados las células marcadas se encontraban dentro de círculos de diámetros de 3, 4, 5, 7, y 8 pixeles.
En este punto es importante aclarar que debido a que las imágenes se consideran como arreglos matriciales, cuando se marcan círculos mediante el software de ImageJ u otro, aunque visualmente parezcan círculos en realidad son una aproximación de la circunferencia con cuadros, esto se ilustra mejor en la Figura (3.3).

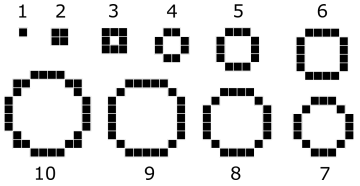

Figura 3.3 Circunferencias de diferentes diámetros formadas con pixeles, a pesar de que con el software Fiji se aprecian círculos en realidad se trabaja con aproximaciones a circunferencias.

Después de que el experto ha marcado las células en un stack el paso siguiente es adquirir las series de tiempo de la variación de la intensidad asociadas a cada célula lo cual se realiza promediando los niveles de intensidad dentro de cada círculo y para cada frame en el stack, obteniendo así un punto en el tiempo con cada frame analizado y que en conjunto conforman la serie de tiempo asociada a la célula dentro del círculo, esto se realiza para todas las células encontradas y se lleva a cabo también con el software ImageJ. Las series de tiempo así encontradas son analizadas posteriormente con el software *Igor Pro* [57] que permite realizar análisis de datos de manera rápida.

Ya que el proceso de la detección de las células es inherentemente humano no está exento de errores y esto se puede ver en la Figura (3.4) donde se muestra de forma ilustrativa un ejemplo del tipo de series de tiempo que se pueden encontrar en los stacks, en este caso las gráficas a, e, f y h no se consideran como asociadas a células por lo que se puede considerar que están mal identificadas. Hay que tomar en cuenta que en un solo stack de imágenes puede haber hasta unas 500 células o más, por esta razón identificarlas manualmente se convierte en una tarea demasiado exhaustiva y que puede provocar errores de detección.

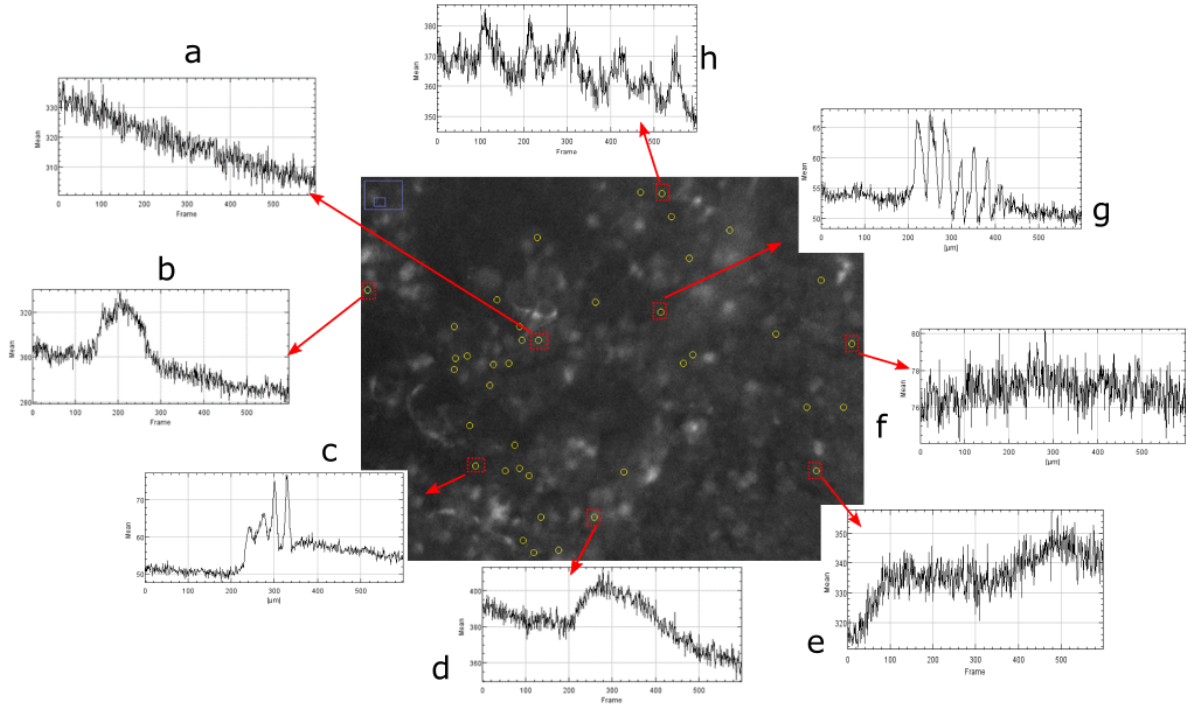

Figura 3.4 Ejemplos de células identificadas en una porción de un stack de imágenes junto con sus series de tiempo asociadas. Nótese el aumento y disminución en la intensidad en las series b, c, d y g, las células asociadas a las series a, e, f y h se considera que están mal identificadas.

Como se mencionó en la introducción en general los trabajos existentes sobre segmentación de células en imágenes basan su funcionamiento en la identificación de los bordes de las mismas [58]–[61], esto se debe a que comúnmente las imágenes poseen un número pequeño de células y muchas veces el mayor problema al que se enfrentan los algoritmos desarrollados para tal fin es la forma irregular de las células o que estén unidas, sin embargo en este trabajo ese no es el problema ya que en realidad la identificación de las células por el experto se basa en sus series de tiempo asociadas porque no se

cuenta con bordes que las definan. Por todo lo anterior el enfoque usado para encontrar las células fue las series de tiempo de las mismas aunque primero se procesaron los stacks para poder trabajar con series de tiempo adecuadas, esto se detalla a continuación.

3.2 Pre-procesamiento de los stacks de imágenes

El primer paso para poder trabajar sobre las series de tiempo fue tratar de remover el decaimiento debido al fotoblanqueamiento, la razón de aplicar esta corrección es porque en el caso ideal sería mucho más sencillo comparar series de tiempo pensando que solo existen dos tipos, uno en el que solo hay ruido y corresponde a una línea casi horizontal (del estilo de la gráfica f en la Figura (3.4)), y otro en el que se presentan uno o más máximos pero también sobre una base casi horizontal, similar a la gráfica g de la Figura (3.4). Claramente es imposible poder llegar al caso ideal debido a que no todas las zonas en el tejido biológico se comportan igual, solo basta con observar la serie de tiempo e de la Figura (3.4) donde no se presenta el decaimiento característico si no hay un aumento en la intensidad, sin embargo se buscó una aproximación que pudiese funcionar para el problema.

Debido a que no es posible hacer una corrección por fotoblanqueamiento sobre las series de tiempo, ya que es lo que se está buscando, el cambio se aplicó sobre los stacks de imágenes para lo cual se observó el comportamiento del histograma a lo largo del tiempo en diferentes stacks, un ejemplo de esto se presentan en la Figura (3.5) en dondes se puede apreciar cómo cambia el histograma a lo largo del tiempo produciendose un efecto de desplazamiento hacia la izquierda, es decir a los niveles de gris más oscuros, mientras que el histograma mantiene una forma muy parecida a lo largo del tiempo.

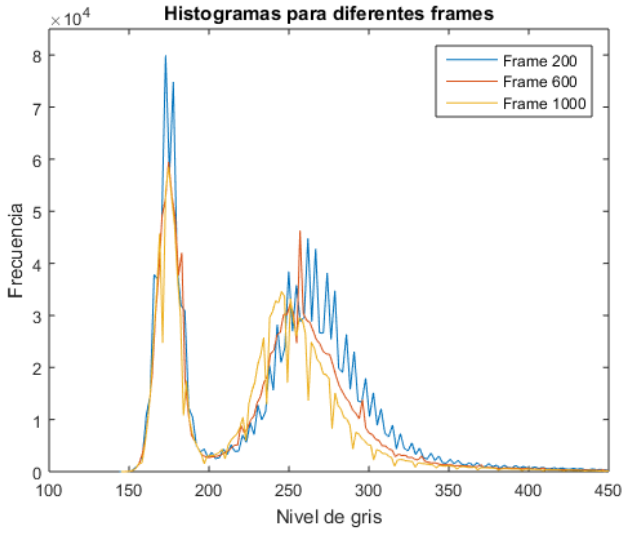

Figura 3.5 Histogramas de diferentes frames para un solo stack, se puede apreciar un corrimiento a la izquierda, es decir a los niveles de gris más oscuros.

Debido a este resultado se propuso implementar la corrección por fotoblanqueamiento usando la especificación del histograma [7], con esto se obligó a que todos los histogramas de cada frame en el stack tuviesen una forma aproximada a la del primer frame, eliminando así el decaimiento, como se mencionó antes esto no dio una solución perfecta, ver Figura (3.6), pero fue bastante útil para poder comparar entre sí las series de tiempo.

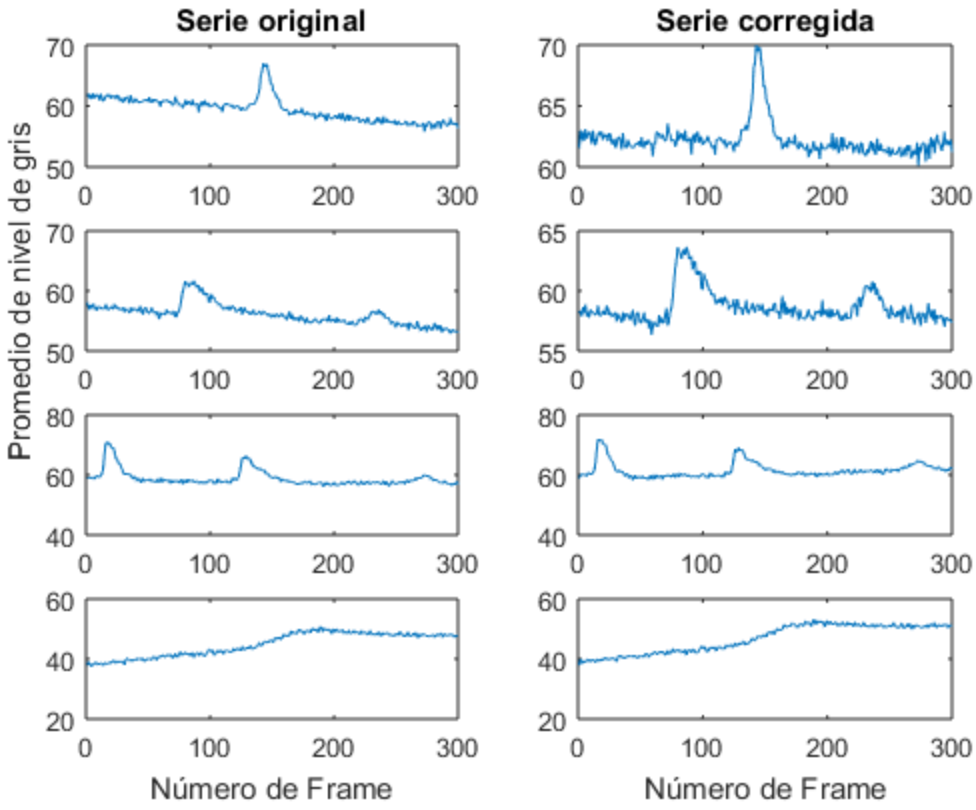

Figura 3.6 Se observan cuatro ejemplos de series de tiempo con y sin corrección por fotoblanqueamiento mediante la especificación del histograma, se puede ver que para las dos primeras series se obtuvieron buenos resultados, para la tercera y cuarta series se aprecia un pequeño levantamiento en ambas en su parte final.

3.3 Selección de características de las series de tiempo

Después de lo que se ha descrito hasta ahora sobre el problema la propuesta para su solución se puede resumir de la siguiente forma, se desea clasificar las series de tiempo asociadas a cada pixel presente en un stack de imágenes en dos clases correspondientes a la presencia o no de una célula, esto se puede traducir en la detección de la existencia o no de uno o más máximos en una serie de tiempo y esto se puede lograr de forma más sencilla cuando se ha removido la tendencia de decaimiento. Es claro que la información proporcionada por un solo pixel no será la misma que la proporcionada por un conjunto de pixeles, recordemos que las series de tiempo se adquieren mediante promedios, pero se espera que con la clasificación emerjan las células como conjuntos de pixeles y que se puedan diferenciar bien de los que no corresponden a células.

Sin embargo el problema de la detección de máximos en series de tiempo se puede volver bastante complicado cuando existe demasiado ruido en la señal, que es el caso de las series de tiempo a analizar, basta con observar las Figuras (3.4) y (3.6), además hay que añadir que el ruido de las series de tiempo se incrementará porque se va a trabajar sobre pixeles y no con conjuntos de ellos. Asimismo para localizar los máximos en una serie de tiempo es de gran ayuda conocer de antemano la frecuencia de oscilación de la señal sin embargo de la Figura (3.4) es fácil entender que conocer esto es complicado debido a que algunas oscilaciones son muy lentas, como en la serie de tiempo d, y otras son muy rápidas, como la serie de tiempo g, dando un intervalo grande de posibles frecuencias.

Debido a que se antoja como algo complicado la localización correcta de máximos en las series de tiempo, en este trabajo se propuso buscar otras características que den información de la presencia o no de máximos de manera indirecta y que se usen para implementar un clasificador de regresión logística, es importante recordar del capítulo 2 que se debe procurar que las características a utilizar sean sencillas de calcular, lo cual cobra gran relevancia en el presente trabajo ya que el clasificador se aplicará sobre $M \times N$ pixeles, donde M y N son el número de renglones y columnas de cada frame en el stack respectivamente, por lo tanto si el stack es demasiado grande y el cálculo de las características excesivamente complejo se traducirá en un gran consumo de tiempo por parte del clasificador.

Inicialmente se eligieron cinco características de estudio asociadas a cada serie de tiempo, el área bajo la curva, el promedio, la desviación estándar, el sesgo y la curtosis (ver el apéndice 3 donde se detalla los momentos estadísticos), aunque solo se deseaba usar a lo más tres de ellas para que resultara sencillo visualizar el espacio de características, si bien las medidas estadísticas de las series de tiempo no dan información suficiente sobre el comportamiento del sistema debido a que son no estacionarias (es decir su media y su varianza cambian en el tiempo), se pudo comprobar que pueden ayudar bastante cuando se usan en el clasificador. La Figura (3.7) da un ejemplo de cómo cambian las distribuciones de probabilidad de algunas series de tiempo junto con sus primeros cuatro momentos y el área.

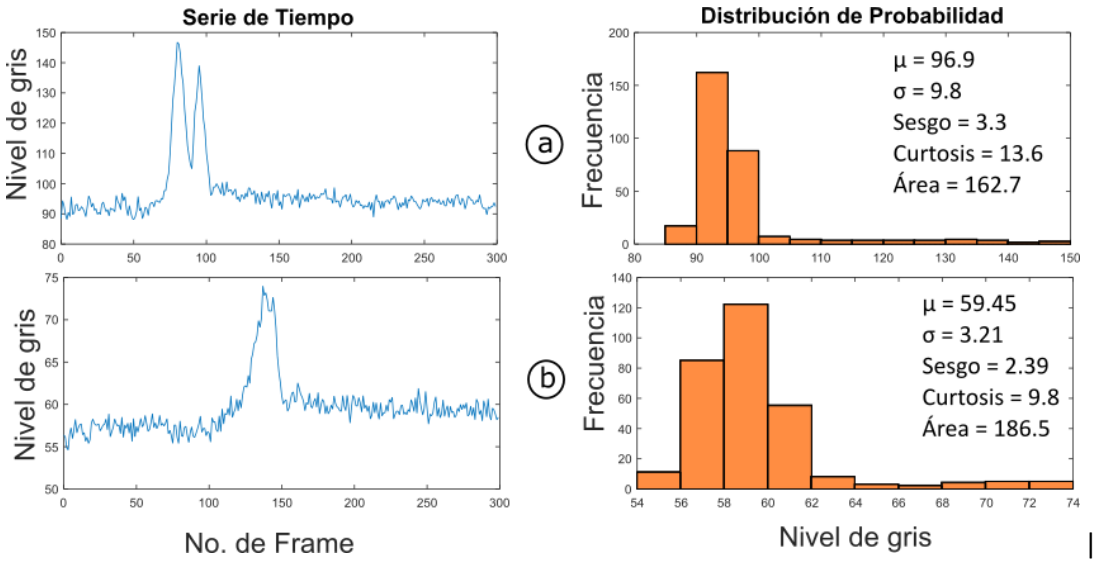

Figura 3.7 Primera parte Continúa en la siguiente hoja

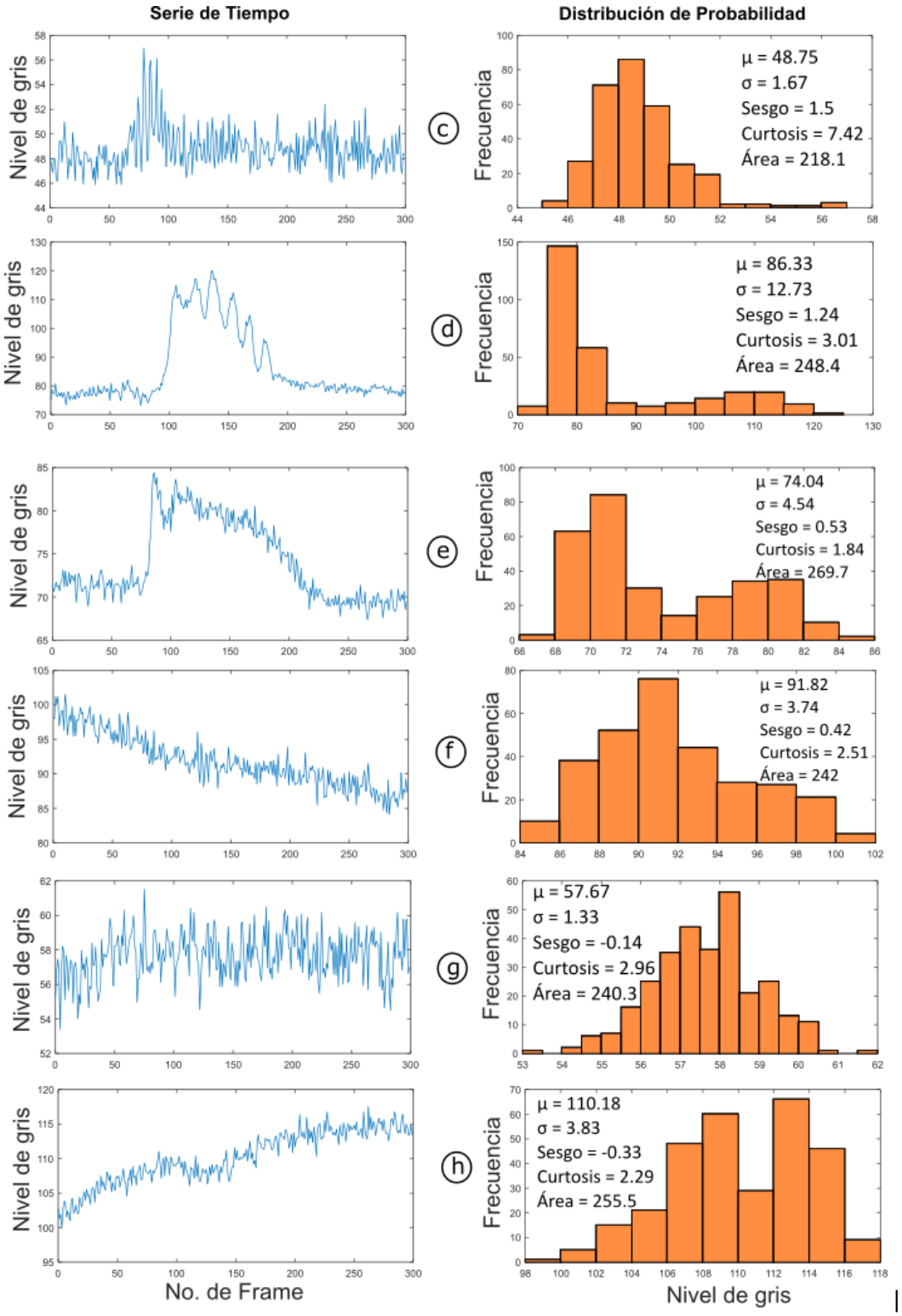

Figura 3.7 Segunda parte Ejemplos de series de tiempo con sus correspondientes histogramas (distribuciones de probabilidad) y los valores de sus características. Las series de tiempo a-e corresponde a una célula, el resto son correspondientes a ruido. Las series de tiempo ya están corregidas por fotoblanqueamiento.

Observando la Figura (3.7), es claro que el promedio no parece dar información relevante relacionada con la forma de la serie de tiempo sobre todo porque presentan niveles de gris muy variados haciendo que su promedio cambie mucho. La desviación estándar tampoco parece ayudar a diferenciarlas, aunque hay una ligera tendencia a que los valores de σ sean más pequeños en los casos de series de tiempo que no corresponden a células no parece que sea información suficientemente útil.

A diferencia de los dos primeros momentos, el sesgo parece que sí puede aportar más información sobre la forma, es fácil observar en la Figura (3.7) que cuando se tiene una distribución de probabilidad sesgada a la derecha (sesgo positivo) corresponde a series de tiempo asociadas a la presencia de una célula (gráficas a-d), por otra parte cuando la densidad de probabilidad se vuelve más simétrica (gráficas f, g) o empieza a presentar un sesgo negativo (gráfica h) corresponden a series de tiempo que no están asociadas a una célula, no obstante es importante destacar que los valores de sesgo entre las gráficas e y f son muy similares entre sí, y la primera corresponde a una célula pero la segunda no.

Respecto a la curtosis, en los primeros gráficos de la Figura (3.7) se presentan valores grandes que van disminuyendo poco a poco hasta alcanzar un mínimo en la gráfica e, donde comienza a aumentar de nuevo pero sin llegar a valores tan grandes como los de los primeros gráficos.

Finalmente respecto al área bajo la curva se puede notar que las series de tiempo que poseen picos estrechos tienen asociadas áreas pequeñas en comparación con otro tipo de series, por esta razón es fácil confundir series como d y f al tomar en cuenta dicha característica.

Aunque este análisis no abarca todas las posibles series de tiempo que se pueden presentar sí da una idea general de qué características podrían ser útiles para el clasificador. Se ha visto que ni el promedio ni la desviación estándar parecen dar información relevante sobre la forma de la serie mientras que el sesgo, la curtosis y el área sí lo hacen pero con algunos problemas con series del estilo d y e de la Figura (3.7) porque podrían confundirse con series de tiempo como f, g y h de la misma figura.

Para el cálculo del área primero se normalizaron las series haciendo que tuviesen promedio cero y desviación estándar uno mediante la siguiente transformación:

$$\hat{x} = \frac{x-\mu}{\sigma}$$

Donde x es la serie de tiempo original, μ y σ el promedio y la desviación estándar respectivamente y $\hat{x}$ la serie de tiempo normalizada, de esta manera las series de tiempo sufrieron un desplazamiento como el mostrado en la Figura (3.8) b, ya que las series así transformadas tienen promedio cero al sumar los niveles de gris de manera directa el resultado es cero, por esta razón se hizo una aproximación burda tomando el valor absoluto de la serie, ver la gráfica c de la Figura (3.8), y luego sumando los niveles de gris.

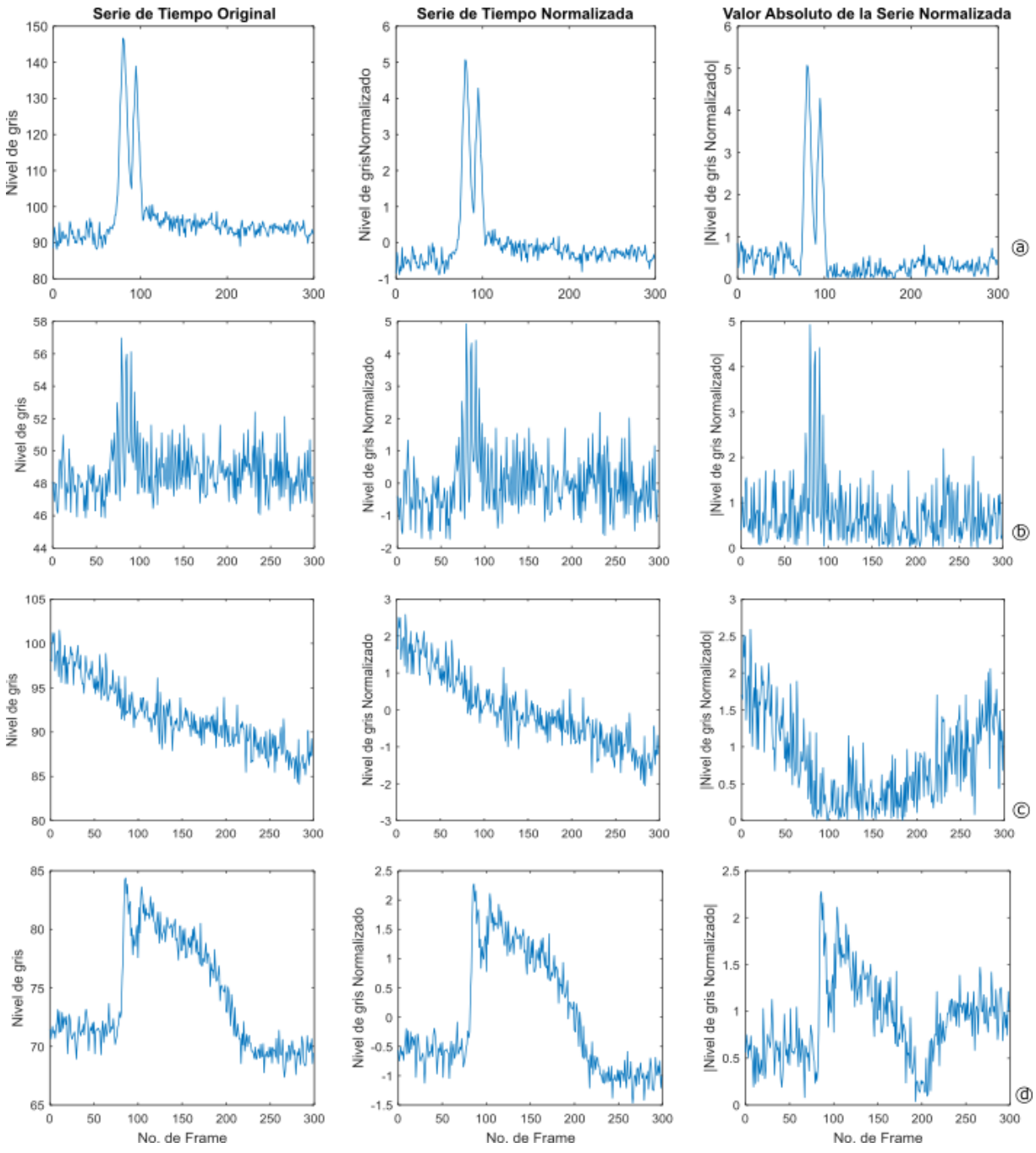

Figura 3.8 Ejemplos de series de tiempo junto con su correspondiente normalización y posterior valor absoluto, esto se hizo con el fin de encontrar un estimado burdo de su área bajo la curva.

Hay que destacar que para poder obtener el área bajo la curva de las series de tiempo fue necesario que todas tuvieran la misma longitud, de no ser así afectaría aún más el cálculo ya de por sí burdo, para esto se modificó el número de frames que poseían los stacks reduciéndolo a 300 (el número de frames más pequeño de todos los stacks con los que se trabajó), para la reducción del número de frames en algunos stacks se eligieron de dos en dos frames, de cuatro en cuatro, etc; hasta alcanzar la cifra deseada,

aunque en otros casos hubo que remover primero una parte del stack debido a que algunos exhiben un incremento en la intensidad generalizado en toda la imagen junto con una posterior disminución debido a la inserción de hormonas como dopamina, tirotropina (TRH) al tejido, etc.

Este cambio generalizado influye enormemente en las series de tiempo, aún cuando se aplica la corrección por especificación del histograma (que ayuda ante cualquier cambio del histograma y no solo al debido por fotoblanqueamiento), ver Figura (3.9), esto también se vio reflejado al usar el clasificador, como se verá más adelante.

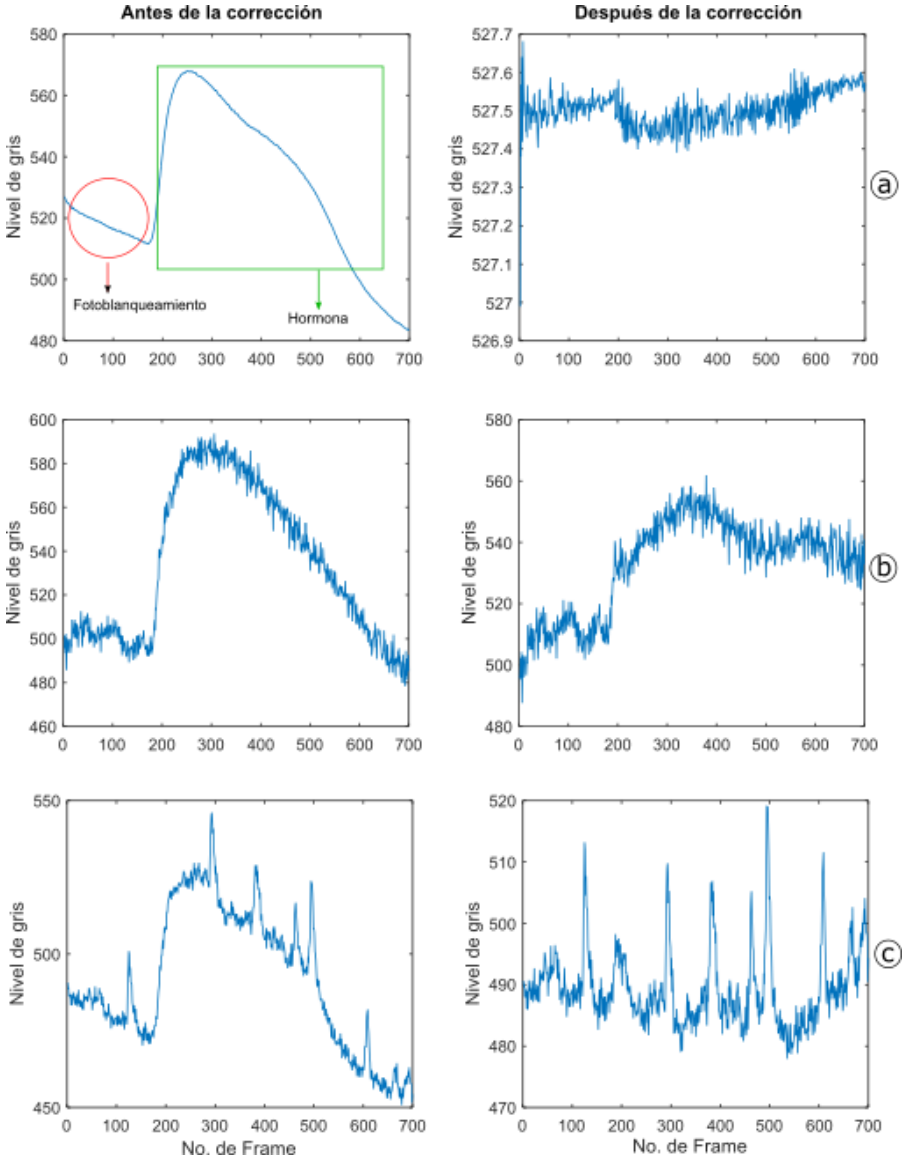

Figura 3.9 a y b muestran el promedio de los niveles de gris por frame de dos diferentes stacks, como se observa además del fotoblanqueamiento existe un cambio añadido debido a la inserción de hormonas en el tejido, del lado derecho se muestra el resultado de corregir el stack con especificación del histograma, ya que el cambio es muy fuerte aún es posible apreciarlo en la corrección. En la gráfica c se observa una serie de tiempo de una célula, y cómo se tienen pequeños picos montados en la señal debida a la inserción de hormonas en el stack, a la derecha se observa esa misma serie de tiempo corregida.

De los 12 stacks que se tenían para trabajar se eligieron 6, de cada uno se seleccionaron 20 series de tiempo para entrenamiento, 10 asociadas a una célula (uno y múltiples máximos) y 10 asociadas a ruido, dando un total de 120 series de tiempo para entrenamiento, de cada serie se obtuvieron cuatro características: desviación estándar, sesgo, curtosis y área bajo la curva, descartando el promedio por lo descrito anteriormente. Finalmente se graficaron los espacios de características tomando en cuenta sólo 2 en cada gráfica y considerando todas las combinaciones sin repetición (resultando ser seis) ya que se deseaba poder apreciar fácilmente su comportamiento, dichas gráficas se muestran en la Figura (3.10), se puede apreciar que no se aplicó normalización sobre las características porque se desconoce cuál será el rango que puedan alcanzar las características para nuevas series de tiempo.

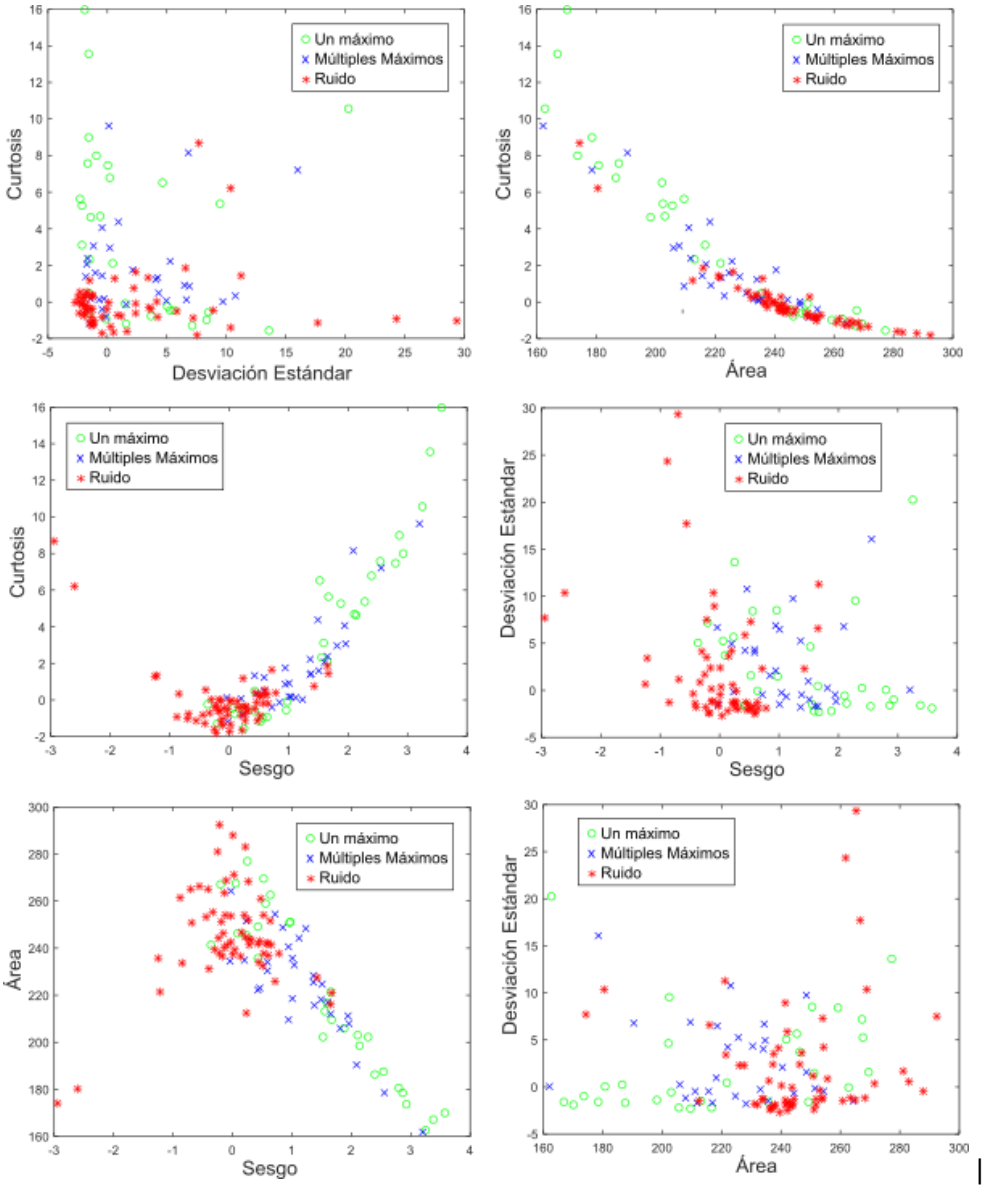

Figura 3.10 Espacios de características de 2D usando el área, desviación estándar, sesgo y curtosis, se marcaron en verde los puntos que corresponden a células con un solo máximo en su serie de tiempo y en azul las que corresponden a más de un máximo en su serie de tiempo.

De la seis gráficas en la Figura (3.10) es posible apreciar que cuando se usa la desviación estándar como característica no hay una separación clara entre los puntos que corresponden a células de los que no, sobre todo en la gráfica f. Por otra parte cuando se usan las otras tres características se aprecia una separación más evidente entre ambas clases aunque en todos los casos existe una superposición parcial de las mismas que será causa de errores de clasificación, sin embargo era un resultado esperado por el análisis realizado a partir de la Figura (3.7).

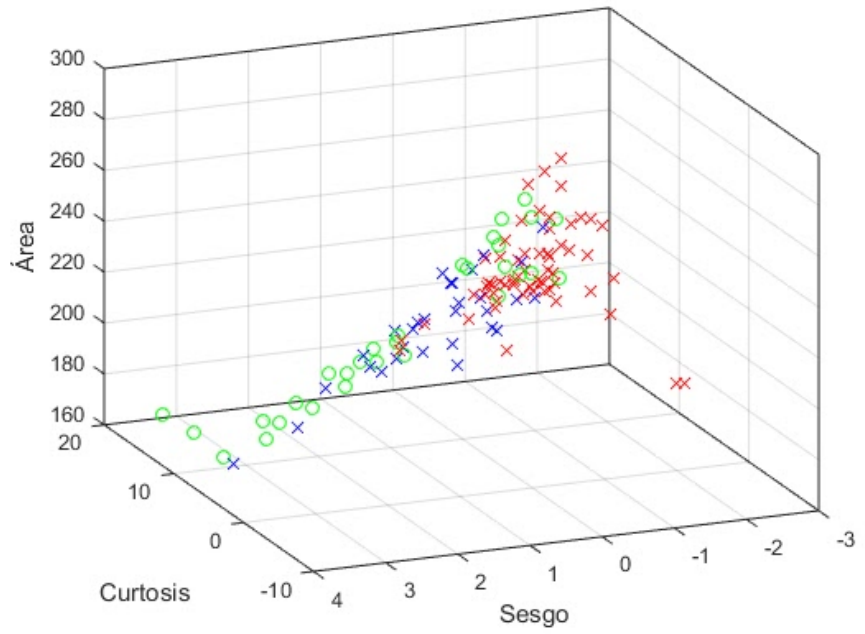

Figura 3.11 Espacio de características en 3D usando el área, sesgo y curtosis, se marcaron en verde los puntos que corresponden a células con un solo máximo en su serie de tiempo y en azul las que corresponden a más de un máximo en su serie de tiempo. Se puede observar que no hay una clara división de las clases en este espacio de características

Para finalizar esta sección se muestra en la Figura (3.11) el caso del espacio de características en tres dimensiones, usando el área, sesgo y curtosis, ya que las gráficas b, c y e son proyecciones de este espacio es claro que ni en él será posible separar por completo las dos clases deseadas.

3.4 Entrenamiento del clasificador y test de funcionamiento

Después de analizar las gráficas de la Figura (3.10) se decidió usar los espacios de características Área-Sesgo y Curtosis-Sesgo para implementar dos clasificadores de regresión logística y comparar su funcionamiento. Inicialmente fue necesario encontrar las fronteras de decisión mediante las ecuaciones (2.10) y (2.11) en su versión extendida ya que ahora se tienen dos características, hay que recordar que en el capítulo 3 se indicó que dichas ecuaciones no se pueden resolver de manera directa dado que no son lineales por esta razón se implementó un algoritmo en el software *Matlab* que usa la función

`fminunc` encargada de encontrar el mínimo de una función no lineal sin restricciones, en este caso sobre β.

Dicha función utiliza un método de cuasi-Newton [62] para encontrar el mínimo por lo que requiere como variables de entrada la ecuación (3.9), extendida para dos características, y el gradiente mediante ecuaciones similares a las presentadas en 3.10 y 3.11, además requiere de un punto inicial a partir del cual comenzar la búsqueda del mínimo. Para verificar que la función converge al mínimo el algoritmo se ejecutó 5 veces para cada clasificador con diferentes puntos de inicio, los resultados obtenidos se muestran en la Figura (3.12) y Tabla (3.1).

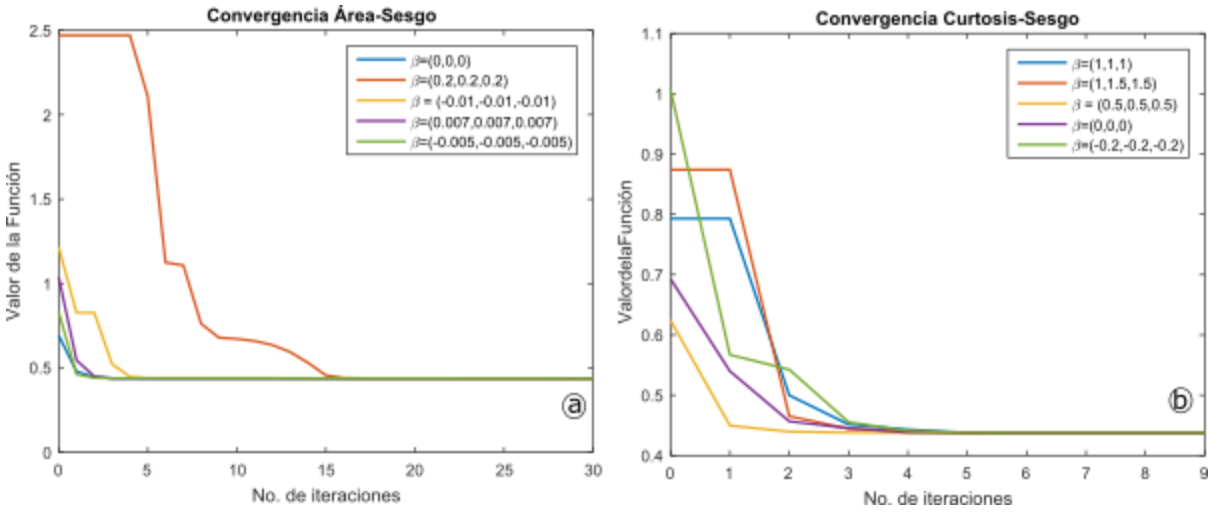

Figura 3.12 Convergencia de los dos clasificadores usando diferentes puntos de inicio mediante la función `fminunc` del software Matlab. Se observa que en todos los casos se converge al mismo resultado en pocas iteraciones.

Como se puede observar en la Figura (3.12) aunque se tengan puntos iniciales diferentes al final todos convergen a un punto similar, en el caso del espacio Área-Sesgo la convergencia se presentó después de 15 iteraciones, mientras que para el espacio Curtosis-Sesgo fue más rápida presentándose después de 5 iteraciones. Por otra parte en la Tabla (3.1) se muestran los valores de β obtenidos para cada punto inicial establecido en las gráficas de la Figura (3.12) cuando se presentó la convergencia, el dato que muestra más variaciones es β_0 para el espacio Área-Sesgo a diferencia del resto que resultaron bastante constantes.

Algo que es notable de la tabla (3.1) es que los valores de β_2 en ambos espacios de características son considerablemente más pequeños que los valores de β_1, los referidos valores de β_2 están asociados a la Curtosis y al Área, c y e de la Figura (3.10) respectivamente, por lo que al final del análisis se puede ver que el Sesgo es la característica que tiene el mayor peso de las tres mientras que el Área es la característica menos sobresaliente.

	Puntos de inicio	β_0	β_1	β_2
Área vs Sesgo	$\beta = (0, 0, 0)$	-2.6	2.3	0.005
	$\beta = (0.2, 0.2, 0.2)$	-2.8	2.4	0.006
	$\beta = (-0.01, -0.01, -0.01)$	-2.5	2.3	0.005
	$\beta = (0.007, 0.007, 0.007)$	-2.9	2.4	0.006
	$\beta = (-0.005, -0.005, -0.005)$	-2.4	2.3	0.004
Promedio		-2.6	2.3	0.005
Curtosis vs Sesgo	$\beta = (1, 1, 1)$	-1.2	2.1	0.1
	$\beta = (1, 1.5, 1.5)$	-1.2	2.0	0.1
	$\beta = (0.5, 0.5, 0.5)$	-1.2	2.1	0.1
	$\beta = (0, 0, 0)$	-1.2	2.1	0.1
	$\beta = (-0.2, -0.2, -0.2)$	-1.2	2.1	0.1
Promedio		-1.2	2.1	0.1

Tabla 3.1 Valores de β obtenidos para los dos clasificadores usando diferentes puntos de inicio en el algoritmo de convergencia y sus promedios.

Las fronteras de decisión se obtuvieron al promediar los valores de β exhibidos en la Tabla (3.1) y se muestran en la Figura (3.13), como se puede observar dichas fronteras se modelaron con un comportamiento lineal ya que se desconoce si para otras series de tiempo las características ocuparán espacio en la zona central de las "parábolas" que se forman.

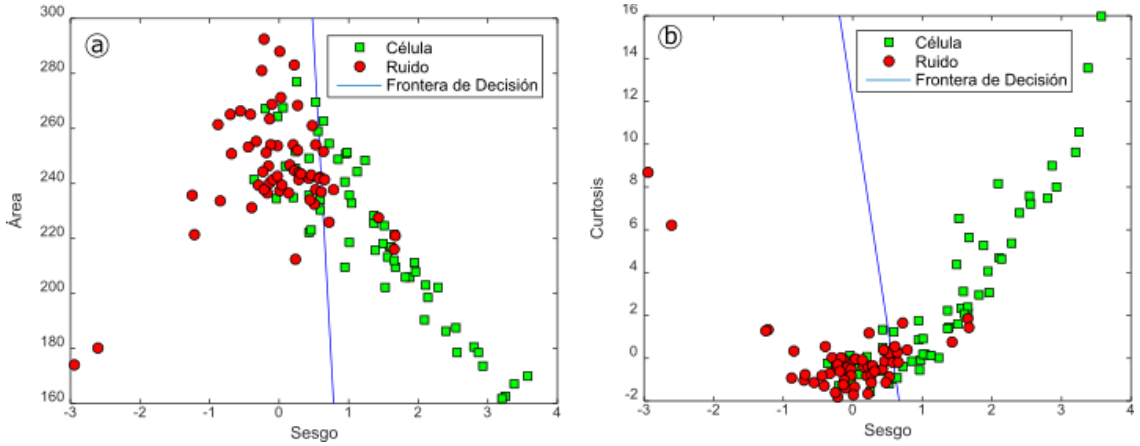

Figura 3.13 Espacios de características junto con las fronteras de decisión lineales encontradas.

Para conocer el error de los clasificadores anteriores se eligieron 60 series de tiempo diferentes a las usadas en el entrenamiento pero de los mismos stacks, 30 series están asociadas a células y 30 corresponden a ruido así que de cada stack se seleccionaron 5 series de células y 5 series de ruido. El resultado para este nuevo conjunto de series se muestra en la Figura (3.14), nótese que aún se siguen preservando las formas de “parábola” que se obtuvieron en la Figura (3.13).

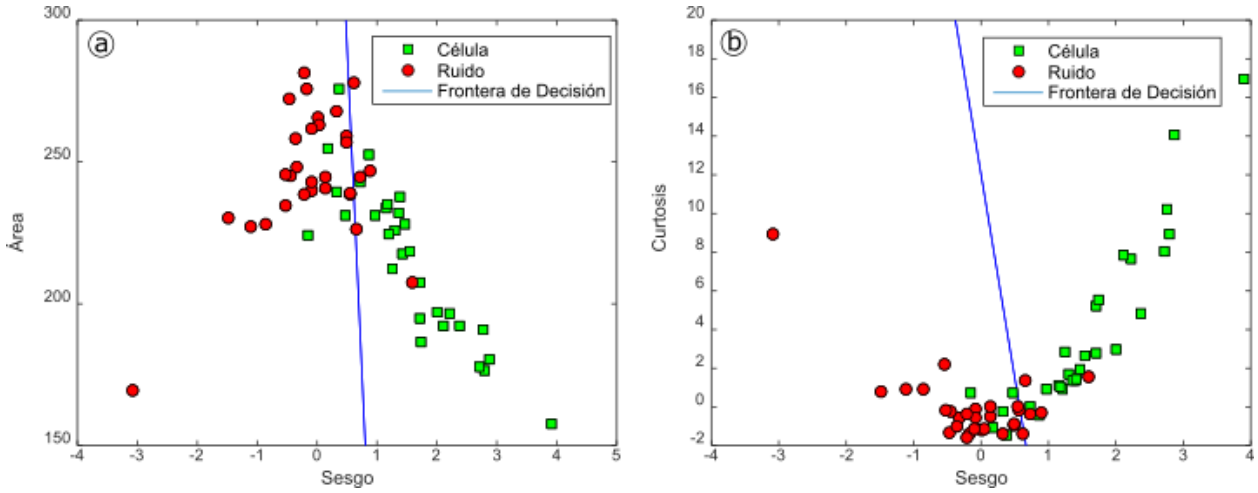

Figura 3.14 Espacio de características con las nuevas series de tiempo y las fronteras de decisión encontradas antes, nótese que las formas son similares a las encontradas en la Figura 3.13 .

Para el caso del espacio Área-Sesgo se presentaron 10 errores de clasificación, lo cual corresponde a un 17% del total de datos aproximadamente, la diferencia con el espacio Curtosis-Sesgo solo fue de un error que corresponde a un 15% así que se puede concluir que ambos clasificadores funcionarán de manera similar, aunque esto se estudia en el siguiente apartado sobre los stacks de imágenes completos.

3.5 Detección de células en stacks de imágenes

Como recapitulación de lo que se ha mencionado hasta ahora, de los 12 stacks de imágenes con los que se contaba al inicio se usaron 6 para extraer 180 series de tiempo, la mitad correspondientes a células y el resto a ruido, dos tercios del total de las series se usaron como conjunto de entrenamiento y el resto en el conjunto de test, cada conjunto posee la mitad de series correspondientes a células y la mitad a ruido. Después de conocer las dos fronteras de decisión de los dos clasificadores fue posible aplicar los clasificadores a los 12 stacks.

Como pre-procesamiento todos los stacks fueron corregidos por fotoblanqueamiento mediante la especificación del histograma, este procedimiento resultó ser el más lento ya que para los stacks más grandes tardó 30 minutos en realizarse en una computadora con 32Gb de memoria RAM usando el software ImageJ. Algunos stacks no se corrigieron adecuadamente ya que presentaban otro comportamiento además del fotoblanqueamiento, ver Figura (3.9).

Finalmente para aplicar los clasificadores a los stacks se desarrolló un programa en Matlab donde cada pixel del primer frame tiene asociada una serie de tiempo formada por los valores de ese pixel en cada frame, y de cada serie de tiempo así obtenida se calcula su área, sesgo y curtosis, es decir las características de los espacios empleados antes, sin embargo ya que antes se ha evaluado el área con series de tiempo de 300 datos fue necesario recortar los stacks que tenían más de 300 frames ya sea removiendo la parte del stack que tiene un comportamiento diferente al del fotoblanqueamiento, haciendo un submuestreo de la serie o aplicando ambos.

Una vez calculados dichos valores fue posible aplicar las fronteras de decisión previas para determinar si el pixel pertenece a la clase célula o a la clase ruido, esta regla se aplica a cada pixel en el primer frame del stack y se genera una imagen del mismo tamaño que la del primer frame donde se usa la codificación pixel blanco = clase célula, pixel negro = clase ruido, es decir se genera una imagen en blanco y negro, donde las aglomeraciones de pixeles blancos se espera indicarán la presencia de una célula en dicho sitio. En la Figura (3.15) se muestran algunos ejemplos de las imágenes así obtenidas.

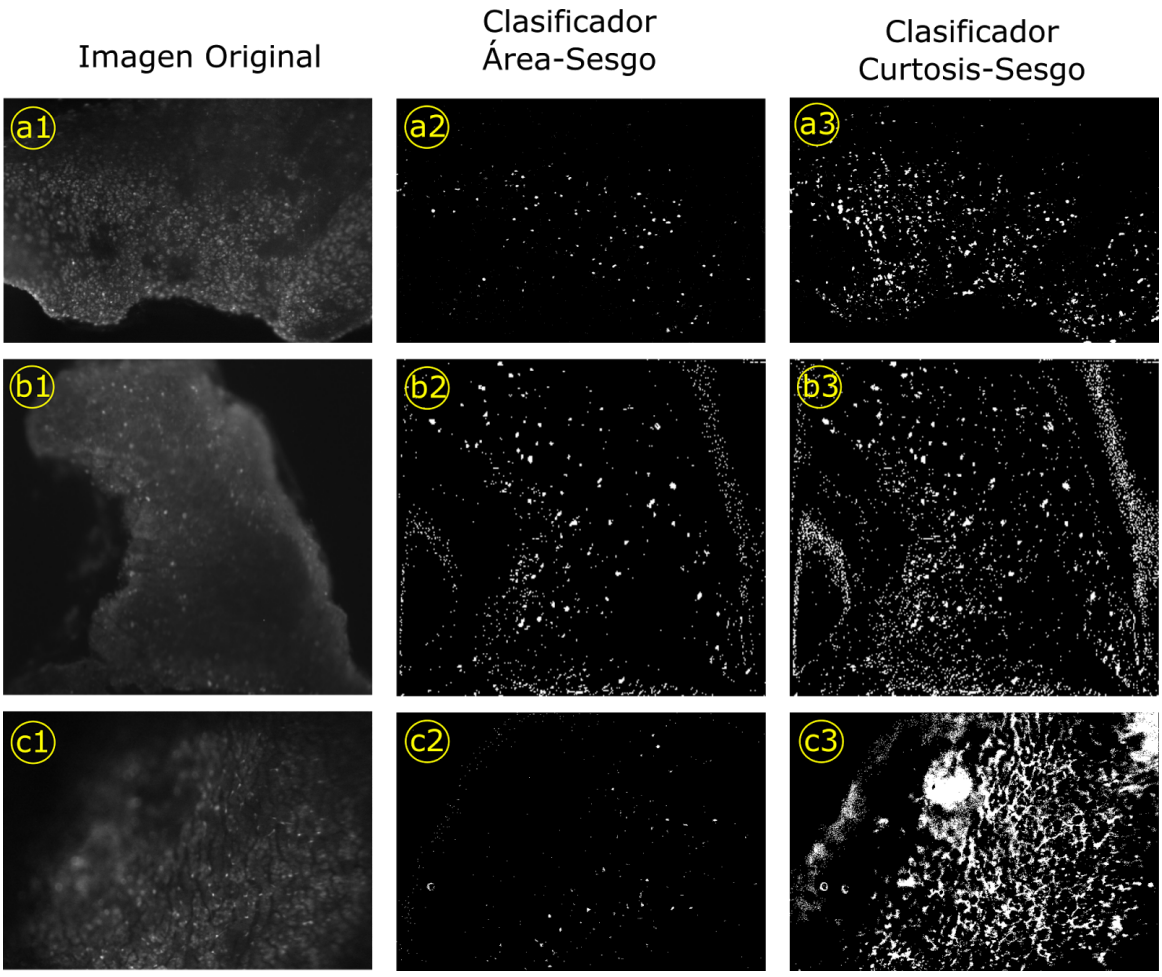

Figura 3.15 Ejemplos del resultado de los dos clasificadores para tres diferentes stacks de imágenes, se espera que las aglomeraciones de pixeles blancos representen una célula. Se puede observar que para el caso del stack c) el clasificador curtosis-sesgo no dio un buen resultado, imagen c3).

Como se puede apreciar de la Figura (3.15) las nuevas imágenes presentan ruido tipo sal que se puede remover fácilmente con un filtro mediana [7]. También se puede observar que el clasificador Curtosis-Sesgo tiene un mayor número de aglomeraciones de pixeles blancos en comparación al clasificador Área-Sesgo, sin embargo también ha aumentado el ruido ya que en la imagen b se puede notar como en zonas donde ya no hay tejido se presentan numerosos pixeles blancos que no pueden pertenecer a células.

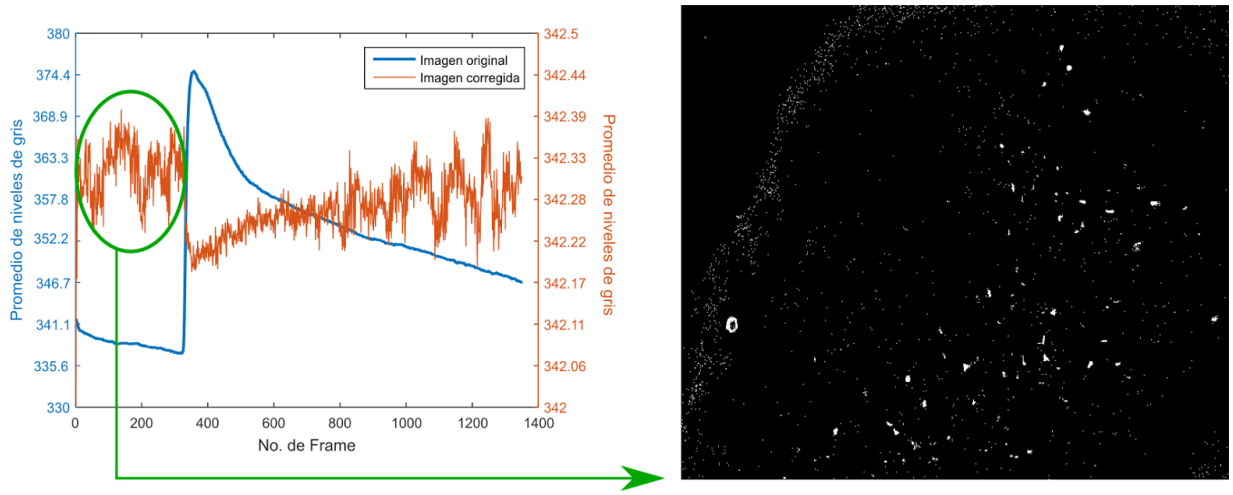

Figura 3.16 El stack c) de la Figura 3.14 presenta otra señal además del fotoblanqueamiento, por eso se tuvo un mal resultado en el clasificador curtosis-sesgo, sin embargo al remover esa zona del stack y aplicar nuevamente el clasificador se tiene un resultado completamente diferente que se asemeja más a lo que se busca.

Asimismo es importante resaltar lo que ocurrió en el caso de la imagen c3 ya que en esta no es posible distinguir acumulaciones separadas de pixeles blancos, esto se debe a que dicha imagen se obtuvo usando el stack corregido cuyo original presentaba fotoblanqueamiento e incremento repentino en la intensidad con su posterior disminución. Lo anterior se comprobó al aplicar ese mismo clasificador a la parte del stack corregido que solo presentaba fotoblanqueamiento, el resultado se muestra en la Figura (3.16), como se puede observar al hacer esto es posible generar una imagen similar a y b de la Figura (3.15), por esta razón los stacks que presentaban el citado comportamiento se analizaron como el stack de la Figura (3.16). En el siguiente apartado se describen los resultados obtenidos mediante los dos clasificadores sobre los 12 stacks.

Resultados

Una vez obtenidas las 24 imágenes con aglomeraciones de pixeles blancos (2 clasificadores por stack), los puntos aislados se removieron con un filtro mediana de 3x3, posteriormente se usó el software Fij para generar contornos que encierran las aglomeraciones y las regiones así encontradas se superpusieron sobre los stacks corregidos por fotoblanqueamiento, Teniendo las regiones superpuestas a los stacks fue posible corroborar manualmente si corresponden a una célula o no, los resultados para cada stack usando los dos clasificadores se resumen en la Tabla (1).

Cabe aclarar que en en algunas imágenes a pesar de aplicarles el filtro mediana se preservaron regiones en zonas donde no existía tejido por lo que se removieron manualmente ya que era obvio que ahí no podían existir células, este efecto se puede ver en la Figura (3.15) en el stack b).

	Clasificador Área-Sesgo			Clasificador Curtosis-Sesgo		
Stack	Regiones halladas	Regiones de ruido	% de error	Regiones halladas	Regiones de ruido	% de error
1	40	4	10%	264	32	12%
2	47	7	14.9%	82	21	25.6%
3	137	12	8.8%	400	8	2%
4	32	25	78.1%	69	57	82.6%
5	87	28	32.2%	149	73	48.9%
6	101	37	36.6%	222	114	51.4%
7	87	24	27.6%	150	71	47.3
8	40	30	75%	109	77	70.6%
9	291	103	35.4%	430	173	40.2%
10	44	8	18.2%	85	15	17.6%
11	22	10	45.5%	139	89	64%
12	114	92	80.7%	142	40	71.8%

Tabla 1 Resultados obtenidos de aplicar los dos clasificadores a los 12 stacks de imágenes.

Para saber si los algoritmos que permiten la detección de células desarrollados en este trabajo pueden proveer de una mejora al actual modo de realizar su detección se consideró el siguiente análisis.

Supongamos que al aplicar los dos clasificadores sobre un stack se obtiene el mínimo número de regiones encontradas mostrado en la Tabla (1), es decir con el clasificador Área-Sesgo se obtienen 22 regiones y con el clasificador Curtosis-Sesgo se obtienen 69, ya que una persona tarda en revisar y corregir 100 regiones en 15 minutos aproximadamente entonces para analizar las regiones del clasificador Área-Sesgo tardará 3.3 minutos y para el clasificador Curtosis-Sesgo tardará 10.4 minutos.

Por otra parte en ese mismo tiempo la persona marcará manualmente 6 y 17 células respectivamente, entonces, para que se pueda admitir que el programa proporciona una mejora se debe cumplir que al menos durante los 3.3 minutos se encontrarán 6 células y durante los 10.4 minutos se encontrarán 17 células, es decir, para el clasificador Área-Sesgo de las 22 regiones encontradas al menos 6 deben corresponder a células y para el clasificador Curtosis-Sesgo de las 69 regiones al menos 17 deben corresponder a células, lo cual representa el 27% y el 25% de acierto respectivamente. Tomando en cuenta lo anterior y considerando además que es preferible encontrar más células que las que manualmente se logran detectar en el mismo tiempo se considerará que el clasificador tiene un buen desempeño cuando tiene a lo más 50% de error.

Considerando esto y de la Tabla (1) resulta que el clasificador Área-Sesgo presentó un buen desempeño en 75% de los stacks mientras que el clasificador Curtosis-Sesgo lo hizo en un 58% de ellos, además los clasificadores presentan un desempeño equiparable al humano en un 75% para el clasificador Área-Sesgo y en un 92% para el clasificador Curtosis-Sesgo, es decir que de aplicarse el clasificador Curtosis-Sesgo sobre un stack es altamente probable que al menos tenga un funcionamiento semejante al desarrollado por una persona en cuanto a número de regiones encontradas y tiempo invertido.

Asimismo, si se promedian los porcentajes de error de los stacks que presentaron un buen desempeño se tiene que el clasificador Área-Sesgo posee un 25.5% de error en promedio, mientras que el clasificador Curtosis-Sesgo muestra un 27.7% de error en promedio, ambos valores se traducen en un 74.5% de aciertos en promedio para el clasificador Área-Sesgo y un 72.3% de acierto para el clasificador Curtosis-Sesgo, sobre las regiones detectadas.

En este análisis no se ha utilizado el hecho de que se requiere de 30 minutos para preprocesar el stack y de 2 minutos más que puede tardar el clasificador en encontrar las regiones, que son los casos extremos cuando los stacks son grandes, debido a que son procesos que hace la computadora sin supervisión y porque es tiempo que el experto puede dedicar para realizar otras actividades, solo se tomó en cuenta el tiempo que el experto realmente requiere destinar para el análisis.

Algo que también se puede observar de analizar la Tabla (1) es que a pesar de que el clasificador Curtosis-Sesgo presenta un mayor error que el de Área-Sesgo, el primero siempre encuentra un mayor número de regiones que sí corresponden a células, sin tomar en cuenta si tiene o no un buen desempeño, lo cual indica que al final sí tuvo un mayor peso la característica de sesgo sobre la del área, justo como se había conjeturado al obtener las ecuaciones de frontera de decisión en el entrenamiento.

De la Tabla (1) se puede observar que los stacks 4, 8 y 12 presentaron un porcentaje de error bastante elevado para los dos clasificadores, en general existen muchas razones por las que no se pudieron

encontrar células adecuadamente, la más obvia de todas ellas es que si se observan las Figuras (3.13) y (3.14) se podrá notar que varias de las series de tiempo que sí fueron clasificadas como células están del lado incorrecto de la frontera de decisión, esto es natural dado que será muy complicado si no es que imposible encontrar un espacio de características que divida en dos clases a los objetos de manera perfecta sin que se presente sobreajuste que produzca peores resultados (ver Capítulo 3), de ahí el error de los clasificadores encontrados.

También recordemos que en una primera instancia dicho error fue de 17% para el clasificador Área-Sesgo y 15% para el de Curtosis-Sesgo sobre el conjunto de test, aunque después se pudo observar que los valores aumentaron bastante cuando se aplicaron sobre stacks completos, esto se debe a que seguramente se encontraron nuevos tipos de series de tiempo que no se tomaron en cuenta durante el entrenamiento, sin embargo es imposible conseguir todos los tipos de series de tiempo para el entrenamiento así que siempre habrá un error en la clasificación aunque se usen otros espacios de características, sólo se puede aspirar a encontrar un espacio que produzca el menor error posible.

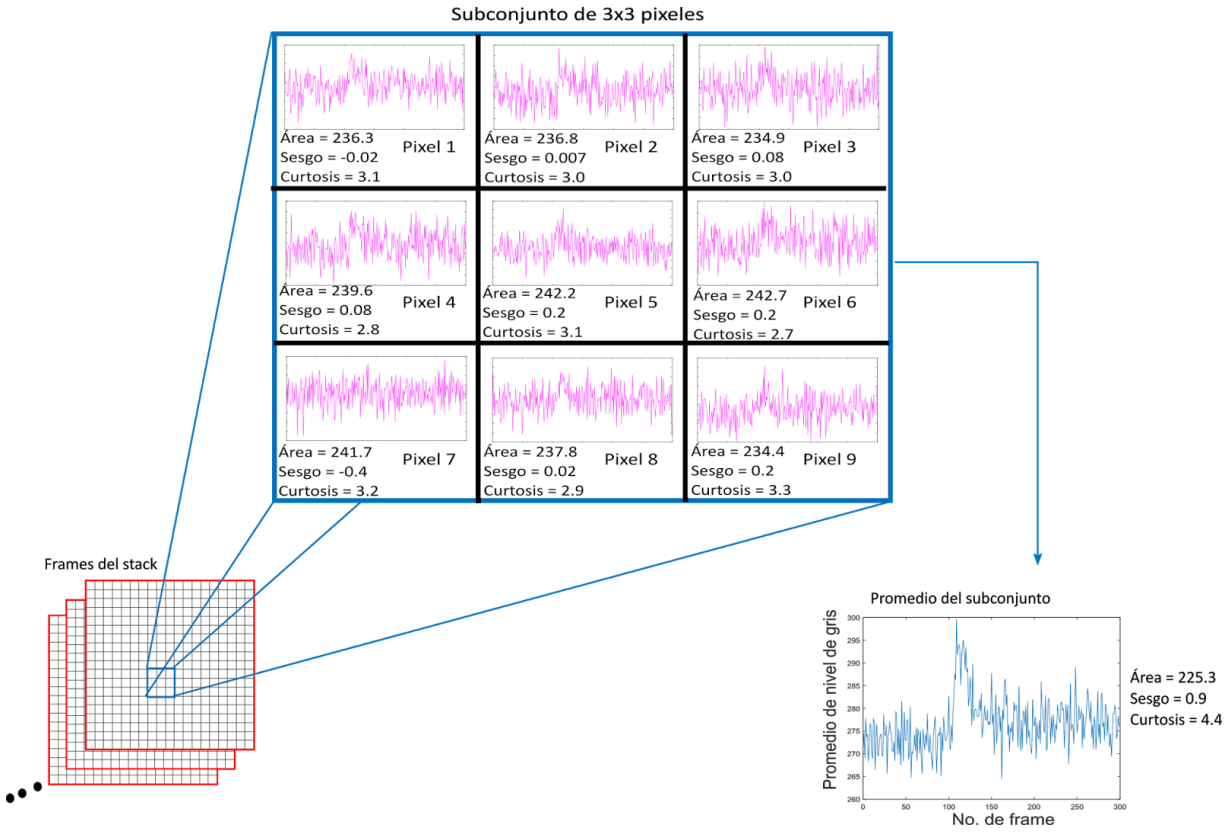

Figura 3.17 Muestra la diferencia entre obtener series de tiempo por pixel dentro de un subconjunto de 3x3 pixeles de un substack y generar la serie de tiempo de ese mismo subconjunto, con los valores obtenidos de área, sesgo y curtosis resulta que mientras las 9 series de tiempo dentro del subconjunto serán clasificadas como no pertenecientes a una célula, la serie de tiempo del subconjunto será clasificada como célula.

Por otra parte hay que resaltar que los clasificadores fueron entrenados sobre series de tiempo que se adquirieron con conjuntos de pixeles, es decir regiones circulares de diámetro 3, 4, 5, 7, y 8 pixeles

dependiendo del stack, sin embargo al aplicarlos sobre los stacks las series de tiempo se obtuvieron por pixel, no por región, con la esperanza de que al generar las imágenes en blanco y negro se tuviesen agrupaciones de pixeles blancos que indicasen la posible presencia de una célula en dicha ubicación, este cambio en el uso de conjuntos de pixeles al de pixeles individuales conducirá a que se tengan zonas en negro (ruido) cuando deberían estar en blanco (célula) como se muestra en la Figura (3.17), donde se muestra que las series de tiempo por pixel indican que los 9 pixeles en el subconjunto no pertenecen a una célula (comparando los valores de área, sesgo y curtosis sobre las fronteras de decisión), sin embargo cuando se toman los 9 pixeles como un todo y se promedian los niveles de gris para formar una nueva serie de tiempo los clasificadores arrojan que esa serie de tiempo sí pertenece a una célula.

Finalmente otra fuente de error está en la reducción simplista del comportamiento de las células en los stacks, a pesar de que muchas células sí poseen el comportamiento de aumento y disminución en la intensidad una o varias veces en el tiempo existen otras cuyo comportamiento está muy lejos de ser así, esto se observó especialmente en el stack 12 de la Tabla (1) ya que a pesar de que a simple vista el stack muestra bastantes células al revisar sus series de tiempo asociadas se generan gráficas como las mostradas en la Figura (3.18), que al final van a ser consideradas como ruido por ambos clasificadores .

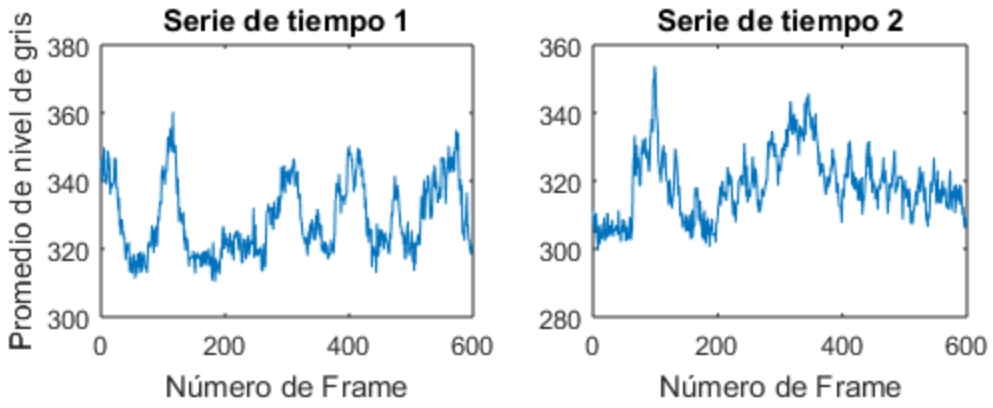

Figura 3.18 Muestra dos series de tiempo asociadas a células obtenidas del stack 12 de la Tabla (1), como se puede observar presentan un comportamiento más complejo que el que se deseaba encontrar con los clasificadores implementados en el trabajo.

Conclusiones y trabajo a futuro

En el presente trabajo se implementó un método de detección semi-automática de células hipofisarias de ratón en stacks de imágenes adquiridas mediante microscopía de fluorescencia. Habitualmente cuando se busca detectar células en imágenes, ellas suelen contener un número reducido de células lo cual permite poder distinguir sus bordes, por esta razón usualmente los problemas a los que se enfrenta la gente que desea detectar células en imágenes es localizar adecuadamente los bordes irregulares de las mismas, usar técnicas de tracking para seguirlas cuando presentan movimiento o incluso lograr encontrar adecuadamente los límites que separan a una célula de otra cuando están demasiado próximas. Sin embargo en este trabajo el principal obstáculo a superar fue la inexistencia de bordes visibles en las células con las que se trabajó y que para su detección el experto requiere analizar su comportamiento de emisión, es decir sus cambios en la intensidad en el tiempo, además de la enorme cantidad de células que pueden estar presentes en un stack.

Ya que el uso de los bordes es inadecuado para el tipo de imágenes con que se trabajó, la búsqueda de las células se basó en las series de tiempo asociadas a las mismas y que se generan de promediar los niveles de gris en las regiones donde se localizan las células para cada frame del stack, localizadas previamente por el experto de manera manual. La justificación para el uso de las series de tiempo se debe a que una gran cantidad de ellas presentan comportamientos específicos cuando sí están asociadas a una célula, es decir aumentos en la intensidad seguidos de una disminución (“picos”), los cuales fueron el fundamento para el empleo de un clasificador como solución del problema planteado.

Se implementaron dos clasificadores de regresión logística cuyas características se eligieron de entre 5 iniciales que se fueron descartando mediante un análisis previo, al final se mantuvieron tres características, área bajo la curva, sesgo y curtosis, con ellas se generaron dos clasificadores Área-Sesgo y Curtosis-Sesgo y se aplicaron sobre 12 stacks de imágenes para conocer su funcionamiento.

Con base en el resultado de los 12 stacks, el desempeño de los clasificadores se evaluó sobre la premisa de que el tiempo que emplea el experto en encontrar un número determinado de células manualmente debe ser igual o mayor al tiempo que requerirá para determinar si la regiones encontradas por el algoritmo son o no células, con la condición de que el clasificador ubique al menos el mismo número de células que encontraría si lo hiciera manualmente. Bajo esta condición se encontró que el clasificador Área-Sesgo se desempeña de forma similar a como lo hace el experto en un 75% de las veces contra un 91% de veces que lo logra el clasificador Curtosis-Sesgo, esto evaluado sobre el número de regiones encontradas por el clasificador no sobre las regiones que el experto puede encontrar manualmente.

Además se encontró que el clasificador Curtosis-Sesgo siempre encuentra un mayor número de regiones candidatas a célula a diferencia del clasificador Área-Sesgo, y aunque con esto el error sobre el primer clasificador también se incrementa al final siempre se tuvieron más regiones correspondientes a células que las encontradas con el segundo clasificador, esto indica que la característica de Área tuvo una menor

contribución que la de la curtosis, además también hizo evidente que la característica con el mayor impacto fue el sesgo.

Con esto se puede concluir que el clasificador Curtosis-Sesgo resultó ser mejor que el clasificador Área-Sesgo, además tiene la ventaja de que se puede aplicar sobre el stack completo, a diferencia del segundo clasificador que requiere de una reducción en su número de frames para poder obtener el valor del área, y tiene una alta probabilidad de que al aplicarse sobre un stack de imágenes al menos tenga un desempeño similar al que tendría una persona de hacer la selección manualmente.

Ya que en la literatura encontrada hasta ahora no existe una solución para el problema abordado en este trabajo se puede afirmar que con este estudio se ha dado un paso importante para poder implementar un algoritmo de detección de células cuyos bordes no están definidos mediante un clasificador de series de tiempo, se encontraron características fáciles de calcular y bastante simples que resultaron ser muy útiles en el proceso de encontrar células dentro de los stacks de imágenes y se ha dado la pauta para que el clasificador Curtosis-Sesgo sea mejorado de tal forma que sea posible encontrar un mayor número de regiones que sí corresponden a células, es decir a disminuir los errores de clasificación.

Además es importante resaltar que se tienen bien identificadas las limitaciones a las que se enfrenta el clasificador, la principal es que siempre existirá un error de clasificación por eso se buscará que la detección sea semi-automática para que el experto pueda determinar si hay regiones mal identificadas o faltantes. Otra limitación es que existen casos donde las células tienen series de tiempo complejas, ver Figura (3.18), por lo que los clasificadores serán incapaces de encontrarlas y si la mayor parte de las células en el stack tienen comportamientos complicados se tendrán resultados bastante malos. La tercer limitación se refiere al caso en que algunos stacks además de fotoblanqueamiento presentan un efecto cambiante generalizado debido a la inserción de hormonas al tejido ver Figura (3.9). Una cuarta limitación se refiere a que el comportamiento de un conjunto de pixeles no será fácilmente reflejado en sus elementos individuales ver Figura (3.17).

Para enfrentar la primera limitación mencionada antes se buscará reunir un mayor número de características adecuadas y realizar un análisis previo para saber de antemano que mejorarán el clasificador Curtosis-Sesgo, ejemplos de estas podrían ser los primeros componentes de la transformada de Fourier, la entropía de las series, su derivada [63], etc.

Además del aumento en el número de características para la mejora del clasificador, en el trabajo a futuro se requiere que todo esté concentrado y unificado dentro de una sola interfaz gráfica, de hecho ya se han realizado avances en esa dirección con el desarrollo de dicha interfaz en Python, sin embargo aún falta trabajo por realizar, la interfaz debe permitir que antes de aplicar la clasificación sobre el stack se puedan remover zonas que no tienen tejido para que el clasificador no explore esas regiones en búsqueda de células, ya que la final se traducirá en tiempo mal empleado, esto se podría realizar de forma semi-automática usando la umbralización del histograma con ayuda del experto.

Es importante destacar que este trabajo está basado en los stacks de imágenes que amablemente facilitó la Dra. Tatiana Fiordelisio Coll encargada del Laboratorio de Neuroendocrinología Comparada de la Facultad de Ciencias de la UNAM y su equipo de trabajo.

Apéndices

Apéndice 1

Se desea expresar de forma diferente la siguiente expresión:

$$L(\overline{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}[1 - \pi_i]^{1-y_i}$$

Aplicando logaritmo en ambos lados de la ecuación:

$$ln[L(\overline{\beta})] = ln\{\prod_{i=1}^{n} \pi_i^{y_i}[1 - \pi_i]^{1-y_i}\} = ln\{\pi_1^{y_1} \cdot \pi_2^{y_2} ... \pi_n^{y_n} \cdot [1 - \pi_1]^{1-y_1} \cdot [1 - \pi_2]^{1-y_2} ... [1 - \pi_n]^{1-y_n}\}$$

Usando la propiedad $ln(a \cdot b) = ln(a) + ln(b)$:

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [ln\{\pi_i^{y_i}\} + ln\{[1 - \pi_i]^{1-y_i}\}]$$

Usando la propiedad $ln(a^b) = b \cdot ln(a)$:

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [y_i \cdot ln\{\pi_i\} + (1 - y_i) \cdot ln\{1 - \pi_i\}]$$

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [y_i \cdot ln\{\pi_i\} - y_i \cdot ln\{1 - \pi_i\} + ln\{1 - \pi_i\}]$$

Usando la propiedad $ln(a/b) = ln(a) - ln(b)$:

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [y_i \cdot ln\{\frac{\pi_i}{1-\pi_i}\} + ln\{1 - \pi_i\}]$$

Tomando en cuenta que $E[y_i] = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} = \pi_i$, se tiene que:

$$1 - \pi_i = 1 - \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} = \frac{1+e^{-\beta_0-\beta_1 x_i}-1}{1+e^{-\beta_0-\beta_1 x_i}} = \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$$

$$\frac{\pi_i}{1-\pi_i} = \frac{1}{\pi_i^{-1}-1} = \frac{1}{e^{-\beta_0-\beta_1 x_i}}$$

Sustituyendo:

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [y_i \cdot ln\{\frac{1}{e^{-\beta_0-\beta_1 x_i}}\} + ln\{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\}]$$

$$ln[L(\overline{\beta})] = \sum_{i=1}^{n} [y_i \cdot \{\beta_0 + \beta_1 x_i\} - ln\{1 + e^{\beta_0+\beta_1 x_i}\}]$$

Apéndice 2

Se desea maximizar la siguiente expresión para β:

$$ln[L(\beta)] = \sum_{i=1}^{n}[y_i \cdot \{\beta_0 + \beta_1 x_i\} - ln\{1 + e^{\beta_0+\beta_1 x_i}\}]$$

Entonces hay que derivar respecto a β_0 y β_1:

$$\frac{\partial ln[L(\beta)]}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}] = \sum_{i=1}^{n}[y_i - \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}] = 0$$

$$\frac{\partial ln[L(\beta)]}{\partial \beta_1} = \sum_{i=1}^{n}[y_i \cdot x_i - \frac{(e^{\beta_0+\beta_1 x_i})\cdot x_i}{1+e^{\beta_0+\beta_1 x_i}}] = \sum_{i=1}^{n} x_i \cdot [y_i - \frac{1}{1+e^{-\beta_0-\beta_1 x_i}}] = 0$$

Estas son las dos ecuaciones que hay que resolver de forma simultánea para β_0 y β_1.

Apéndice 3

Es posible obtener información cuantitativa de los histogramas (y por ende de la función de densidad de probabilidad) a través de los momentos estadísticos y las medidas de tendencia central, permitiendo así realizar comparaciones y conocer el grado de similitud o diferencia entre diferentes funciones de densidad de probabilidad. Algunas de estas medidas se describen a continuación.

Media aritmética o promedio

Es el primer momento y el más conocido, describe alrededor de qué valor tienden a agruparse un conjunto de mediciones por lo que también es una medida de tendencia central. Para un conjunto de n datos $\{x_1, x_2, x_3, ..., x_n\}$, el promedio se define como:

$$\mu(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Varianza

Es el segundo momento y uno de los más conocidos, describe qué tan dispersos están los datos alrededor de la media. Para un conjunto de n datos $\{x_1, x_2, x_3, ..., x_n\}$ con media μ la varianza se encuentra mediante:

$$\sigma^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu(x))^2$$

A partir de este momento se define la desviación estándar $\sigma(x)$ como la raíz cuadrada de la varianza, se utiliza más comúnmente que la varianza y prácticamente da la misma información que ella.

Sesgo

Es el tercer momento, indica qué tan asimétrica es una distribución con respecto a la media, esto se muestra en la Figura (A.1), como se puede observar se dice que una distribución está sesgada a la derecha cuando tiene una cola más larga a la derecha y viceversa.

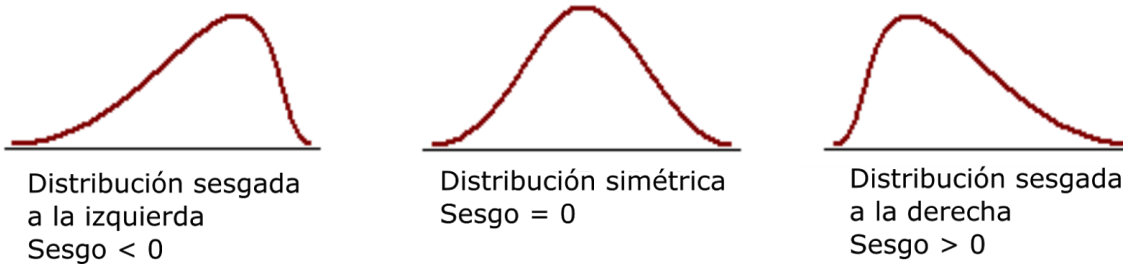

Figura A.1 Ejemplos de distribuciones de probabilidad sesgadas. Se observa que el sesgo es negativo cuando se forma una cola hacia la izquierda y es positivo cuando se forma una cola a la derecha. Tomada de [64].

Para un conjunto de n datos $\{x_1, x_2, x_3, ..., x_n\}$ con media μ y desviación estándar σ el sesgo se obtiene mediante:

$$sesgo = \frac{1}{n\cdot\sigma^3}\sum_{i=1}^{n}(x_i - \mu(x))^3$$

Este coeficiente es adimensional y será positivo si la distribución está sesgada a la derecha, negativo si está sesgada a la izquierda y cero si es simétrica.

Curtosis

Es el cuarto momento, indica que tan elevada o aplanada está una distribución respecto a la distribución normal, esto se muestra en la Figura (A.2), se dice que la distribución es platicúrtica cuando es aplanada, mesocúrtica cuando se trata de la distribución normal y leptocúrtica cuando la distribución es elevada.

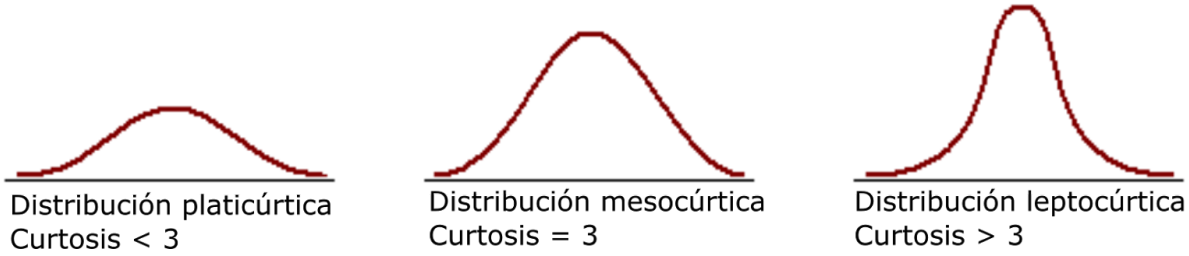

Figura A.2 Ejemplos de distribuciones de probabilidad con diferentes valores de curtosis. Se observa que para un valor de curtosis menor a 3 la distribución es aplanada y para un valor de curtosis mayor a 3 es elevada, mientras que al tener una curtosis = 3 corresponde a una distribución normal. Tomada de [64].

Para un conjunto de n datos $\{x_1, x_2, x_3, ..., x_n\}$ con media μ y desviación estándar σ la curtosis se obtiene mediante:

$$curtosis = \frac{1}{n \cdot \sigma^4} \sum_{i=1}^{n} (x_i - \mu(x))^4$$

Este coeficiente es adimensional y será mayor que 3 si la distribución es leptocúrtica, menor que 3 si la distribución es platicúrtica y tres si es mesocúrtica.

Referencias

[1] San Roque, L., Kendrick, K. H., *et al.*, "Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies," *Cognitive Linguistics*. 2015 Feb; 26(1): 31–60.

[2] "MIT Research - Brain Processing of Visual Information," *MIT News on campus and around the world*, 1996 Dic 19;. [Online]: http://news.mit.edu/1996/visualprocessing.

[3] DiCarlo, J. J., Zoccolan, D., & Rust, N. C., "How does the brain solve visual object recognition?," *Neuron*. 2012 Feb 9; 73(3): 415–434.

[4] Cobos, J. Á., "La historia del microscopio," *Revista de Divulgación Científica y Tecnológica de la Universidad Veracruzana*. 2012 Ene 1; XXV(1). [Online]: https://www.uv.mx/cienciahombre/revistae/vol25num1/articulos/historia/

[5] Petrou, M. & Petrou, C., *Image processing: the fundamentals*. 2009, 2da ed. Singapore: John Wiley & Sons.

[6] Sadeghian, F., Seman, Z., *et al*., "A framework for white blood cell segmentation in microscopic blood images using digital image processing.," *Biological procedures online*. 2009 Jun; 11(1):196-206.

[7] Gonzalez, R. C. & Woods, R. E., *Digital Image Processing*. 2002, 2da ed. USA: Pearson Education.

[8] Maška, M., Ulman, V., *et al.*, "A benchmark for comparison of cell tracking algorithms," *Bioinformatics*. 2014 Feb 12; 30(11):1609-1617.

[9] Yang, F., Mackey, M. A., *et al*., "Cell segmentation, tracking, and mitosis detection using temporal context," *Medical image computing and computer-assisted intervention - MICCAI 2005*. 2005; 302–309.

[10] Meijering, E., "Cell segmentation: 50 years down the road [Life Sciences]," *IEEE Signal Processing Magazine.* 2012 Ago 22; 29(5):140-145.

[11] Sonka, M., Hlavac, V., & Boyle, R., *Image processing, analysis, and machine vision*. 2008, 3a ed. USA:International Thomson.

[12] Tata, J. R., "One hundred years of hormones.," *EMBO reports*. 2005 Jun 1; 6(6): 490–496.

[13] Starling, E. H., "The Croonian Lectures on the Chemical Correlation of the Functions of the Body," *The Lancet.* 1905 Ago 5; 166(4275):339-341.

[14] Gorbman, A. & Bern, H. A., *A Textbook of Comparative Endocrinology*. 1962*,* Wiley.

[15] Lechan, R. M. & Toni, R., "Functional anatomy of the hypothalamus and pituitary," *Endotext*. 2016 Nov 28; [Online]: https://www.ncbi.nlm.nih.gov/books/NBK279126/

[16] Melmed, S. & Conn, P. M., *Endocrinology Basic and Clinical Principles*. 2005, 2da ed. USA: Humana Press.

[17] Fauquier, T., Guérineau, N. C., *et al*., "Folliculostellate cell network: A route for long-distance communication in the anterior pituitary," *Proceedings of the National Academy of Sciences*. 2001 May 9; 98(15):8891–8896.

[18] Uhlhaas, P. J., Haenschel. C., *et al*., "The role of oscillations and synchrony in cortical networks and their putative relevance for the pathophysiology of schizophrenia," *Schizophrenia bulletin.* 2008 Jun

17; 34(5):927–943.

[19] Stephens, T. D., Seeley, R. R., & Tate, P., *Essentials of Anatomy & Physiology*. 2001, 4ta ed. McGraw Hill.

[20] Tortora, G. J. & Derrickson, B., *Introduction to the human body, the essentials of anatomy and physiology*. 2010, 8va ed. USA: John Wiley and Sons.

[21] Barret, K. E., Barman, S. M., *et al*., *Ganong Fisiología Médica*. 2010 23va ed. China: McGraw-Hill Interamericana Editores.

[22] Gardner, D. G., & Shoback, D., *Greenspan's: Basic and Clinical Endocrinology*. 2010, 9na ed. China: McGraw Hill.

[23] Fauquier, T., Lacampagne, A., *et al.*, "Hidden face of the anterior pituitary," *Trends in endocrinology and metabolism: TEM*. 2002 Sep 1; 13(7):304–309.

[24] Hodson, D. J., Molino, F., *et al*., "Investigating and modelling pituitary endocrine network function," *Journal of neuroendocrinology*. 2010 Jul 29; 22(12):1217–1225.

[25] Hodson, D. J., *et al.*, "Coordination of calcium signals by pituitary endocrine cells in situ.," *Cell calcium*. 2012 Abr; 51(3–4):222–230.

[26] van Vreeswijk, C., & Sompolinsky, H., "Chaos in neuronal networks with balanced excitatory and inhibitory activity," *Science*. 1996 Dic 16; 274(5293):1724–1726.

[27] Onesto, V., *et al*., "Information in a Network of Neuronal Cells: Effect of Cell Density and Short-Term Depression," *BioMed research international.* 2016 May 10; 2016:1-12.

[28] Patel, T. P., *et al*., "Automated quantification of neuronal networks and single-cell calcium dynamics using calcium imaging," *Journal of neuroscience methods.* 2015 Mar 30; 243:26–38.

[29] Wu, Q., Merchan, F, & Castleman, K. R., 2008, *Microscope image processing*. USA: Academic Press, Elsevier.

[30] Lichtman, J. W. & Conchello, J.-A., "Fluorescence microscopy," *Nature methods*. 2005 Nov 18; 2(12): 910–919.

[31] Yeung, C.-M., *et al*., "Cells of the anterior pituitary," *The international journal of biochemistry & cell biology*. 2006 Mar 03; 38(9):1441–1449.

[32] "ProLong Live Antifade Reagent Protects Fluorescent Proteins and Dyes In Live-Cell Imaging," *Thermo Fisher Scientific*. [Online]: https://www.thermofisher.com/mx/es/home/references/newsletters-and-journals/bioprobes-journal-of-cell-biology-applications/bioprobes-72/bioprobes-72-prolong-live-antifade.html

[33] Hoebe, R. A., *et al*., "Controlled light-exposure microscopy reduces photobleaching and phototoxicity in fluorescence live-cell imaging," *Nature biotechnology*. 2007 Ene 21; 25(2):249–253.

[34] Tsie, R. Y. & Waggoner, A., "Fluorophores for Confocal Microscopy: Photophysics and Photochemistry," *Handbook of Biological Confocal Microscopy*. 1995 3ra ed. USA: Springer.

[35] Dickinson, M. E. & Davidson, M. W., " Introduction to Spectral Imaging and Linear Unmixing," *ZEISS We make it visible: Education in Microscopy and Digital Imaging*. [Online]: http://zeiss-campus.magnet.fsu.edu/articles/spectralimaging/introduction.html

[36] Waters, J. C., "Accuracy and precision in quantitative fluorescence microscopy," *The Journal of cell biology*. 2009 Jun 29; 185(7):1135–1148.

[37] Costa, L. da F. & Cesar, R. M., *Shape Analysis and Classification: Theory and Practice*. 2000, 2da ed.

USA:CRC Press.

[38] Spruyt, V., "The Curse of Dimensionality in classification," Computer vision for dummies. 2014 Abr 16. [Online]:http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

[39] Kotsiantis, S. B., "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*. 2007 Oct., 31:249–268.

[40] Dougherty, E. R., *Probability and Statistics for the Engineering, Computing and Physical Sciences*. 1990. USA: Prentice Hall.

[41] Ng, A. Y. & Jordan, M. I., "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*. 2002; 2:841–848.

[42] Kutner, M. H. *et al*., *Applied Linear Statistical Models.* 2004, 5ta ed. USA:McGraw-Hill/Irwin.

[43] Collet, D., *Modelling Binary Data.* 2003, 2da ed. USA: CRC Press.

[44] Bulmer, M. G., *Principles of Statistics*. 1978 2da ed. USA: Dover Publication.

[45] Ekstrom, C. T. & Sørensen, H., *Introduction to Statistical Data Analysis for the Life Sciences*. 2009. USA: CRC Press.

[46] Hosmer, D. W. & Lemeshow, S., *Applied Logistic Regression*. 2000, 2da ed. USA:John Wiley & Sons.

[47] "Función de distribución," *Wikipedia*, 2017 Abr 01. [Online]: https://es.wikipedia.org/wiki/Funci%C3%B3n_de_distribuci%C3%B3n

[48] Kleinbaum, D. G. & Klein, M., *Logistic regression: a self-learning text*. 2008. USA: Springer.

[49] Mccrea, N., "An Introduction to Machine Learning Theory and Its Applications: A Visual Tutorial with Examples," *Toptal LLC*. [Online]:https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer

[50] Ripley, B. D., *Pattern Recognition and Neural Networks*. 1995. United Kingdom: Cambridge University Press.

[51] Dollár, P. *et al*., "Feature mining for image classification," *Computer Vision and Pattern Recognition CVPR'07. IEEE Conference on*. 2007 Jul 16; 1–8.

[52] Alelyani, S., Tang, J. & Liu, H., "Feature Selection for Clustering: A Review," *Data Clustering: Algorithms and Applications*. 2013. USA: Chapman and Hall/CRC.

[53] Jain, A. & Zongker, D., "Feature selection: Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*. 1997 Feb. 19(2):153–158.

[54] "Introduction to Machine Learning with Naive Bayes," *Tom Robertshaw*. 2015 Dic 21. [Online]: https://tomrobertshaw.net/2015/12/introduction-to-machine-learning-with-naive-bayes/

[55] "Learning Curve," *Ritchie Ng*. 2017 May 29. [Online]: http://www.ritchieng.com/machinelearning-learning-curve/

[56] "Welcome," *ImageJ*. 2016 Sep 20. [Online]: https://imagej.net/Welcome

[57] "Igor Pro 7," *WaveMetrics*. [Online]: https://www.wavemetrics.com/

[58] Leskó, M. *et al*., "Live cell segmentation in fluorescence microscopy via Graph Cut," *Pattern Recognition (ICPR), 2010 20th International Conference on*. 2010 Oct 07. 1485-1488.

[59] Du, X. & Dua, S., "Segmentation of fluorescence microscopy cell images using unsupervised mining,"

The Open Medical Informatics Journal. 2009 Nov 15. 4:41-49.

[60] Dimopoulos, S. *et al*., "Accurate cell segmentation in microscopy images using membrane patterns," *Bioinformatics*. 2014 May 21. 30(18):2644-2651.

[61] Wang, Z. Z., "A New Approach for Segmentation and Quantification of Cells or Nanoparticles," *IEEE Transactions on Industrial Informatics*. 2016 Mar 14. 12(3):962-971.

[62] Baldick, R., *Applied Optimization: Formulation and Algorithms for Engineering Systems*. 2006. USA:Cambridge University Press.

[63] Górecki, T. & Łuczak, M., "Using derivatives in time series classification," *Data Mining and Knowledge Discovery*. 2012 Feb 01. 26(2):310-331.

[64] "Importance of data distribution in training machine learning models," *Coding algorithms*, 2015 Nov 13. [Online]: https://tekmarathon.com/2015/11/13/importance-of-data-distribution-in-training-machine-learning-models/