PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

MODELO COMPUTACIONAL PARA LA IDENTIFICACIÓN DE CÉLULAS ESPERMÁTICAS MEDIANTE EL ANÁLISIS AUTOMÁTICO DE MICROGRAFÍAS DIGITALES

Tesis para optar por el Título de Ingeniero Informático, que presenta el bachiller:

Heidy Hernández Bretón

ASESOR: Dr. César Armando Beltrán Castañón

Lima, Febrero del 2015

Resumen

El presente proyecto de fin de carrera consiste en el desarrollo de un modelo computacional para la identificación de células espermáticas con el objetivo de analizar la normalidad de la morfología de la cabeza de dichas células mediante el análisis de micrografías digitales. El modelo propuesto comprende el procesamiento de las imágenes microscópicas, la extracción y selección de características que identifican la cabeza de las células espermáticas, la clasificación de las mismas en normales o anormales atendiendo a criterios morfológicos y el análisis comparativo de la caracterización realizada con relación a los estándares de la Organización Mundial de la Salud.

Las imágenes microscópicas fueron procesadas para obtener una máscara binarizada de las mismas donde se identificara la cabeza de las células. Posteriormente las cabezas de las células fueron caracterizadas de manera automática de acuerdo a métricas seleccionadas y se realizó una reducción de dimensionalidad utilizando Análisis de Componentes Principales. Para la clasificación se emplearon Máquinas de Soporte Vectorial.

Como resultado del procedimiento aplicado se pudieron identificar el 91.5% de las células espermáticas existentes en las imágenes de muestra. La tasa de acierto conseguida para la clasificación morfológica fue del 77.6%. Las métricas consideradas en la caracterización están de acuerdo a los parámetros de la Organización Mundial de la Salud.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO: Modelo computacional para la identificación de células espermáticas mediante el análisis automático de micrografías digitales.

ÁREA: Ciencias de la Computación
PROPONENTE: César A. Beltrán Castañón
ASESOR: César A. Beltrán Castañón
ALUMNO: Heidy Hernández Bretón
CÓDIGO: 20114419
TEMA N°: 592
FECHA: San Miguel, 14 de diciembre de 2014

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Facultad de Ciencias e Ingeniería
08 ABR. 2015
MIGUEL MEJIA PUENTE
DECANO

DESCRIPCIÓN

El proceso manual que se sigue para el diagnóstico de la infertilidad masculina es una preocupación de gran importancia por su impacto psicológico, social y económico. El análisis de las células espermáticas es el primer paso en la obtención de este diagnóstico. Dicho análisis considera concentración, movilidad y clasificación morfológica, siendo esta última uno de los mejores discriminantes del potencial fértil en humanos.

La Organización Mundial de la Salud (OMS) dictamina un conjunto de valores referenciales para la clasificación de las células espermáticas de acuerdo a su morfología teniendo en cuenta diversas características como la forma de la cabeza de la célula espermática, la forma de la cola, etc.

La revisión del estado del arte hace notar el gran interés existente en la comunidad científica, por tratar de resolver el problema planteado buscando definir métodos objetivos y estándarizados, que permitan facilitar la tarea de la clasificación de células espermáticas, para disminuir la incertidumbre en el diagnóstico de la infertilidad masculina.

Un consenso común es la necesidad de estandarizar los procedimientos y prácticas en todo el proceso de clasificación, desde la manipulación de las muestras, la segmentación, las mediciones, etc. La solución que se propone en el presente trabajo, seguirá las recomendaciones propuestas en estudios anteriores, buscando mejorar los resultados obtenidos previamente en cuanto al porcentaje de acierto conseguido. Para ello, se realizarán mejoras en los procesos de extracción de características de las células, así como en la selección del algoritmo de aprendizaje automático y la calibración de sus respectivos parámetros.

Av. Universitaria 1801
San Miguel, Lima – Perú
Apartado Postal 1761
Lima 100 – Perú
Teléfono:
(511) 626 2000 Anexo 4801

OBJETIVO GENERAL

Desarrollar un modelo computacional que permita caracterizar los aspectos morfológicos de la cabeza de las células espermáticas mediante el procesamiento de micrografías digitales de muestras de esperma y la posterior clasificación de dichas células en normales o anormales mediante técnicas de aprendizaje automático.

OBJETIVOS ESPECÍFICOS

Los Objetivos Específicos (OE) del Proyecto son:

OE1: Procesar las microimágenes para extraer características morfológicas de la cabeza de las células espermáticas.

OE2: Seleccionar las mejores características para realizar la clasificación de las células usando métodos automatizados y analíticos.

OE3: Aplicar un algoritmo de Máquinas de Soporte Vectorial (SVM) a los datos obtenidos en (OE2) calibrando adecuadamente sus parámetros y evaluando los resultados obtenidos.

OE4: Reportar estadísticas sobre la caracterización morfológica de la cabeza de las células y su comparación con los criterios de la Organización Mundial de la Salud.

ALCANCE

El presente proyecto consiste en el desarrollo de un modelo computacional para la caracterización y clasificación de células espermáticas, sobre la base del análisis de micrografías digitales. Para ello se emplearán técnicas de procesamiento de imágenes, reconocimiento de patrones, y aprendizaje de máquina.

Se procesarán las micrografías para segmentar las cabezas de las células espermáticas, posteriormente se extraerán las características y se realizará reducción de las dimensiones utilizando Análisis de Componentes Principales. Luego se procederá a la clasificación de las cabezas de las células en normales o anormales utilizando un algoritmo de Máquina de Soporte Vectorial, validando los resultados obtenidos mediante validación cruzada y matrices de confusión.

Finalmente se elaborará un reporte estadístico sobre los rangos de las características observadas en las células de acuerdo a los resultados de la clasificación, y se compararán dichos rangos con los sugeridos por la Organización Mundial de la Salud para la distinción entre células normales y anormales.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú
Apartado Postal 1761
Lima 100 – Perú
Teléfono:
(511) 626 2000 Anexo 4801

A mis padres, por el esfuerzo que han puesto siempre en mi educación y en mi formación como persona. A mis padres, por todo su amor.

Heidy

AGRADECIMIENTOS

A mi asesor César Beltrán por su apoyo y su motivación en la realización de este proyecto. Sus aportes durante el proceso investigativo fueron cruciales para la obtención de los resultados que se presentan.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

1. INTRODUCCIÓN

En este primer capítulo se expone la problemática que se aborda en el presente trabajo de investigación. También se presenta el estado del arte así como los objetivos que se persiguen y los resultados esperados. El capítulo finaliza con un mapa que muestra la organización del documento.

1.1. Problemática

Según la Organización Mundial de la Salud (OMS) [World Health Organization, 2009], la infertilidad es una enfermedad del sistema reproductivo definida como la incapacidad de lograr un embarazo clínico después de doce meses o más de relaciones sexuales no protegidas. Este problema tiene un amplio impacto psicológico, social y económico. El diagnóstico de la infertilidad es fundamental para detectar las causas de la enfermedad, especificar un tratamiento en los casos en que esta sea reversible, o dar la opción a parejas de optar por una variante de reproducción asistida donde esta sea irreversible.

La Sociedad Americana para la Medicina Reproductiva [American Society for Reproductive Medicine, 2006] plantea que el análisis de las células espermáticas es el primer paso y el más importante en la evaluación de la infertilidad masculina. Dicha evaluación también incluye un examen físico, evaluación hormonal, análisis genético, entre otros. El análisis de las células espermáticas generalmente considera concentración, movilidad y clasificación morfológica, siendo esta última, uno de los mejores discriminadores del potencial fértil en humanos.

En el contexto del análisis de la fertilidad masculina, la clasificación morfológica se refiere a la determinación de si una célula espermática es normal o no teniendo en cuenta su forma. Aquellas células clasificadas como normales deben ser las que tengan potencial de fertilizar el óvulo y en función de esto se definen los criterios de lo que es una “forma normal”. En el capítulo 2 se presenta detalladamente la fundamentación biológica del problema de la clasificación morfológica de células espermáticas.

La principal limitante cuando se habla de análisis morfológico de las células espermáticas, es la gran variedad de formas que pueden adoptar dichas células, de modo que existe una gran variabilidad y subjetividad en los resultados que dan los distintos laboratorios. Por ejemplo, en un estudio realizado a 243 hombres cuyas

esposas estaban embarazadas, se encontró que la media de células espermáticas normales era solamente del 20% [Chia, Tay, & Lim, 1998].

Existen dos métodos para examinar la morfología de la esperma de humanos, ambos basados en el análisis de muestras: el primero es mediante observación visual (método manual), y el segundo es usando herramientas computarizadas [Auger, 2010]. En el caso del método manual existen varios factores que influencian los resultados de los exámenes: las técnicas de tintura y fijación, los procedimientos de manipulación de la esperma, pero sobre todo, las habilidades del evaluador. Como consecuencia de esto, para una misma muestra, puede existir gran variación entre los diagnósticos de diferentes observadores [World Health Organization, 2010]. Esto demuestra la importancia de encontrar métodos más objetivos para resolver este problema.

El análisis automático de la morfología de la esperma (ASMA), tiene el potencial de solucionar muchos de los inconvenientes que resultan del modo convencional de análisis antes descrito, con la ventaja adicional de que puedan detectarse características que no pueden ser identificadas de manera visual [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]. Haciendo uso del método estándar, el observador puede clasificar las células en normales o anormales y, en algunas ocasiones, detectar algunos tipos de anomalías, de acuerdo a criterios previamente definidos. En cambio, las tecnologías asistidas por computador miden diferentes características morfológicas de las células espermáticas, pudiendo alcanzar un mayor nivel de detalle en la clasificación [Auger, 2010].

La OMS ha publicado manuales con el objetivo de estandarizar los procedimientos de análisis de células espermáticas a nivel mundial [World Health Organization, 2010]. Muchos laboratorios han adoptado estos estándares y valores de referencia, pero existe una preocupación generalizada, sobre si los valores propuestos por la OMS son demasiado estrictos, llegando a afirmar que estos debieran ser reconsiderados [American Society for Reproductive Medicine, 2006]. En el año 2010, la OMS publicó la quinta edición de dicho manual [Cooper & Noonan, 2010], donde se perfeccionaron los métodos de medición y se amplió la población de estudio, especificándose nuevos límites más flexibles que los de ediciones anteriores. Sin embargo, estos estudios son todavía limitados en cuanto a la población analizada y los métodos utilizados [Esteves, et al., 2012].

Adicionalmente, la clasificación visual de células espermáticas está sujeta a diversas limitaciones. Existe una gran variabilidad en la diferenciación de células normales y anormales, siendo esta variabilidad mucho mayor cuando se trata de identificar tipos de anomalías, lo cual está relacionado a la continuidad de la forma y tamaño de las distintas partes que componen las células espermáticas [Auger, 2010]. Otro factor de gran importancia es el hecho de que esta clasificación visual está en función de los mecanismos de visión humanos y su integración con el cerebro, el cual es una herramienta poderosa cuando se trata de identificar patrones, pero no tanto si se trata de hacer mediciones [Auger, 2010]. Una forma de mitigar estas limitaciones es el uso de métodos asistidos por computadora [Auger, 2010].

Las herramientas para análisis de esperma asistido por computadora (CASA) han evolucionado por aproximadamente 40 años y son usadas comúnmente en laboratorios clínicos en todo el mundo [Amann & Waberski, 2014]. Sin embargo, los técnicos de estos laboratorios, generalmente están poco capacitados en los principios de funcionamiento de estas herramientas, lo que hace que muchas veces se comentan errores en los diagnósticos, al no poder controlar algunas de las fuentes de error como las derivadas de los instrumentos de medición empleados [Amann & Waberski, 2014]. Las principales críticas a estas herramientas giran alrededor de su alto costo, la dificultad de calibración, validación de resultados y estandarización de parámetros [Pepper-Yowell, 2011].

Las limitaciones tanto del método manual (o visual) como del método automático para la clasificación morfológica de las células espermáticas muestran que todavía existe margen en esta área de investigación para mejorar los resultados obtenidos y aumentar la fiabilidad en las herramientas automatizadas para el diagnóstico de la infertilidad masculina.

1.2. Estado del arte

La presente revisión explora algunas de las soluciones al problema de la clasificación morfológica de células espermáticas dentro del marco del análisis de fertilidad. Los trabajos que se presentan difieren en cuanto a la metodología empleada y a los resultados obtenidos. En la presente sección se introducen algunos términos especializados propios del problema del análisis de fertilidad masculina, en el capítulo 2 se profundiza en el significado de dichos términos.

1.2.1. Aplicaciones de herramientas CASA

Según [Lu, Huang, & Lu, 2013], las primeras referencias del uso del microscopio para el análisis de células espermáticas data del año 1678, cuando el científico holandés Anton van Leeuwenhoek inventó el microscopio óptico, sin embargo, no es hasta el año 1985 que están disponibles las herramientas CASA (computer-aided semen analysis). La esencia de un sistema CASA es proyectar imágenes sucesivas de células espermáticas en un arreglo detector, que detecta objetos basado en la intensidad de los pixeles y usar un software especial para extraer la información deseada y proveer determinados resultados. La sensibilidad y la precisión de dichos resultados, dependen del software utilizado [Amann & Waberski, 2014]. Este tipo de herramientas han sido adoptadas en gran cantidad de clínicas alrededor del mundo. Una encuesta tomada en el año 2010, muestra que aproximadamente el 50% de los laboratorios clínicos usan estas herramientas.

Los primeros sistemas CASA necesitaban de gran cantidad de intervención humana, con una tendencia a la disminución de la misma. Se espera que en el futuro solo sea necesario un operador que se asegure del correcto funcionamiento de la herramienta, que coloque las muestras, comience el proceso y analice los resultados. En este sentido, versiones recientes de CASA tienen incorporadas funciones de control de calidad para los procedimientos de análisis de movilidad, concentración y morfología [Lu, Huang, & Lu, 2013].

A pesar de ser ampliamente usadas, las herramientas CASA tienen varias limitaciones, como por ejemplo, la falta de una definición adecuada de valores de referencia para los parámetros del sistema o cómo determinar la proporción adecuada de células con morfología y movilidad normales en una muestra [Lu, Huang, & Lu, 2013]. En la actualidad hay más de doce sistemas CASA en el mercado usados para el análisis de la esperma de animales, la mayoría de ellos basados en la movilidad de las células [Amann & Waberski, 2014].

1.2.2. Análisis de células espermáticas caninas con el analizador de Hamilton-Thorne

De manera similar a como ocurre con los humanos, las anomalías morfológicas en las células espermáticas de caninos, equinos, bovinos y porcinos, también han sido asociadas a bajas tasas de fertilidad [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]. Un grupo de científicos del Departamento de Reproducción, Obstetricia y Salud

de la Manada de la Facultad de Medicina Veterinaria de la Universidad de Ghent en Bélgica, han experimentado con el analizador Hamilton-Thorne el análisis morfológico de las células espermáticas en perros [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004].

El analizador de Hamilton-Thorne provee información métrica detallada, pero para la obtención de buenos resultados, se hace énfasis en la validación y estandarización de los procedimientos, como por ejemplo, la preparación del semen, el método de tintura, el nivel de magnificación del microscopio y la concentración de las muestras [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004].

El nombre del software es Metrix Oval Head Morphology, y se encuentra implementado en el sistema CEROS de Hamilton-Thorne [Hamilton-Thorne, 2014]. Este sistema consiste en un microscopio óptico Olympus, una cámara, un digitalizador de imágenes, una computadora para guardar y analizar los datos recogidos además de un filtro verde para aumentar el contraste entre las células y el fondo [Hamilton-Thorne, 2014].

En la experimentación con la herramienta, se realizaron las siguientes mediciones de las células espermáticas[1] [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]:

- Longitud de la cabeza (μm)
- Ancho de la cabeza (μm)
- Área total de la cabeza (μm^2)
- Alargamiento (%)
- Perímetro de la cabeza (μm)
- Longitud de la cola (μm)
- Porcentaje de células espermáticas normales (%)

Para que una célula sea clasificada como normal, debe tener todas las métricas mencionadas anteriormente en los intervalos normales. Estos intervalos se observan en la Tabla 1.1 y fueron determinados en estudios preliminares. Después de comparados los resultados obtenidos con la métrica antes mencionada, las muestras serán clasificadas en normales, anormales o rechazadas.

[1] Las partes que componen a la célula espermática y sus métricas asociadas se explican con detalle en el capítulo 2.

Tabla 1.1 Métrica HTR 12.1 para definir los intervalos normales de las características de las células espermáticas. Adaptación de [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004].

Parameter	Normal Range
Major (µm)	5.7 – 7.5
Minor (µm)	3.5 – 4.4
Area (μm^2)	16.2 – 24.5
Elon (%)	47.6 – 67.0
Perimeter (µm)	15.6 – 19.0

Entre los experimentos realizados en este estudio, existe uno que compara los resultados obtenidos con la métrica propuesta y los obtenidos de manera convencional por el método manual. Existe una alta correlación entre estos resultados (Figura 1.1), aunque en el caso del análisis automático se detectó un menor porcentaje de células normales debido a anomalías que no son fáciles de detectar por la visión humana [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004].

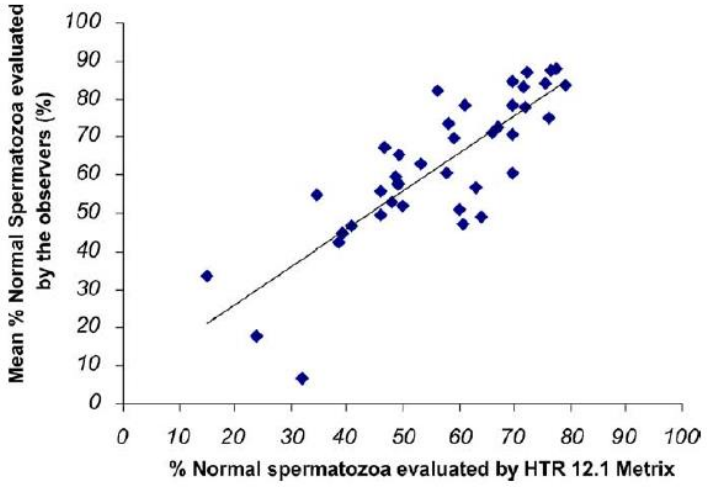

Figura 1.1 Relación entre el porcentaje de células normales detectadas con la métrica HTR 12.1 y mediante observadores humanos [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004].

1.2.3. Analizador automático de la morfología de células espermáticas

Motivados por las limitaciones del método manual para el análisis de células espermáticas, un grupo de profesores de la Universidad del Norte, en Barranquilla, Colombia, desarrolló un método objetivo para el análisis morfológico, teniendo en cuenta el largo y ancho de la cabeza de las células espermáticas, el área de la cabeza, el

porcentaje ocupado por acrosoma, entre otras características [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007].

El método empleado consta de cinco fases:

- Adquisición de imágenes. Muestras preparadas de acuerdo al manual de la OMS [World Health Organization, 2010].
- Detección y extracción de células espermáticas individuales.
- Mejoramiento de la imagen. El objetivo de esta fase es mejorar la definición de las partes que constituyen las células espermáticas identificadas en la parte anterior: núcleo, acrosoma, etc.
- Algoritmo de Segmentación. El objetivo de este algoritmo es subdividir la célula espermática en acrosoma, núcleo y pieza media[2].
- Extracción de características y clasificación. La extracción de las características se realiza sobre las partes definidas en la etapa de segmentación. Estas características serán usadas en la evaluación de la morfología de las células espermáticas. El primer paso es el mapeo entre pixeles y micrómetros. El factor de conversión usado en la presente investigación fue de $0.0288\ pixel/micrometros$.

Las características medidas en la fase de extracción son la longitud de la cabeza, el ancho de la cabeza, el perímetro, el área de la cabeza, el porcentaje de acrosoma, el ancho de la pieza media, entre otros[3]. Una vez que se encuentran los registros de todas estas características en las muestras analizadas, se procede a clasificar las células en normales o anormales de acuerdo al criterio establecido por la OMS, en su cuarta edición [World Health Organization, 1999].

El procedimiento anterior fue probado con una base de datos de 216 imágenes, mostrando excelentes resultados en cuanto a la clasificación de las células espermáticas. Los investigadores plantean la necesidad de validar el método presentado con muestras mayores y con otras especies de mamíferos [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007].

[2] El acrosoma, el núcleo y la pieza media son componentes de las células espermáticas, más información se encuentra en el capítulo 2.

[3] Todos estos términos se explican en detalle en el capítulo 2.

1.2.4. Analizador automático de la morfología de la cabeza de las células espermáticas usando software libre

Un grupo de investigadores se dieron a la tarea de desarrollar una herramienta de código abierto, como alternativa a las caras implementaciones que existen en el mercado. Este trabajo tuvo como objetivo presentar una solución de análisis automático de morfología de células espermáticas, en forma de un plug-in para ImageJ, que es un sistema de código abierto para el procesamiento y análisis de imágenes [Butts, et al., 2011]. La validez del mismo se comprobó usando muestras de semen de bacalao (especie de pez).

El procedimiento utilizado consistió en capturar y medir 30 muestras de esperma, usando el plug-in implementado (Figura 1.2), comparar el ancho y largo de la cabeza medidos por el plug-in con los medidos de manera manual, determinar la precisión de la herramienta usando formas de cabezas previamente calibradas y establecer los parámetros de la detección de imágenes del plug-in usando diferentes valores de contraste y brillo [Butts, et al., 2011]. Como resultado, se comprobó la validez de esta herramienta, como una alternativa a otras implementaciones existentes más costosas [Butts, et al., 2011].

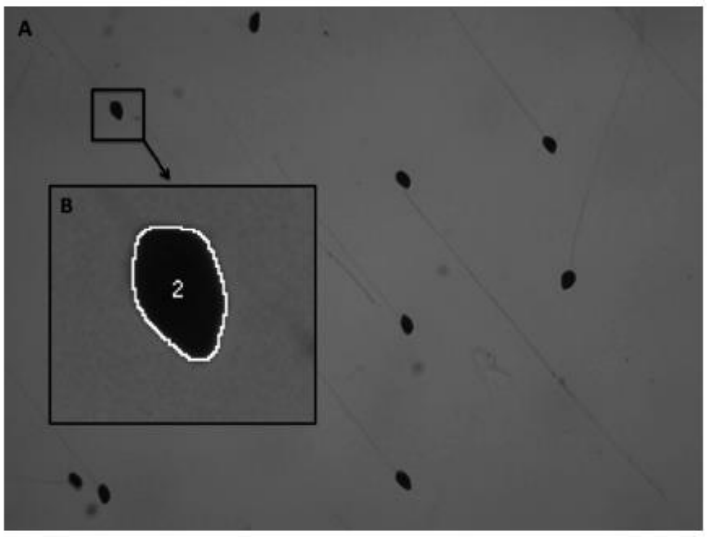

Figura 1.2 Imagen de la esperma del pez bacalao, antes (A) y después (B) de la aplicación de la herramienta de análisis automático [Butts, et al., 2011].

1.2.5. Máquinas de soporte vectorial aplicadas al diagnóstico de la morfología de las células espermáticas

Para resolver el problema del análisis morfológico de células espermáticas mediante la clasificación de imágenes, se han analizado varios acercamientos. El estudio que se expone a continuación, detalla en el método algorítmico usado para lograr este objetivo: máquinas de soporte vectorial [Tseng, et al., 2013].

El sistema está equipado por un microscopio conectado a una computadora para observar imágenes de células espermáticas en tiempo real. Una vez extraída la imagen, se realiza una clasificación con máquinas de soporte vectorial. Finalmente, se realiza una comparación entre este método y otros métodos previamente usados, como el k-ésimo vecino más cercano, el modelo elíptico, y el SIFT (Scale-Invariant Feature Transform) [Tseng, et al., 2013]. El procedimiento se muestra resumido en el diagrama de flujo de la Figura 1.3.

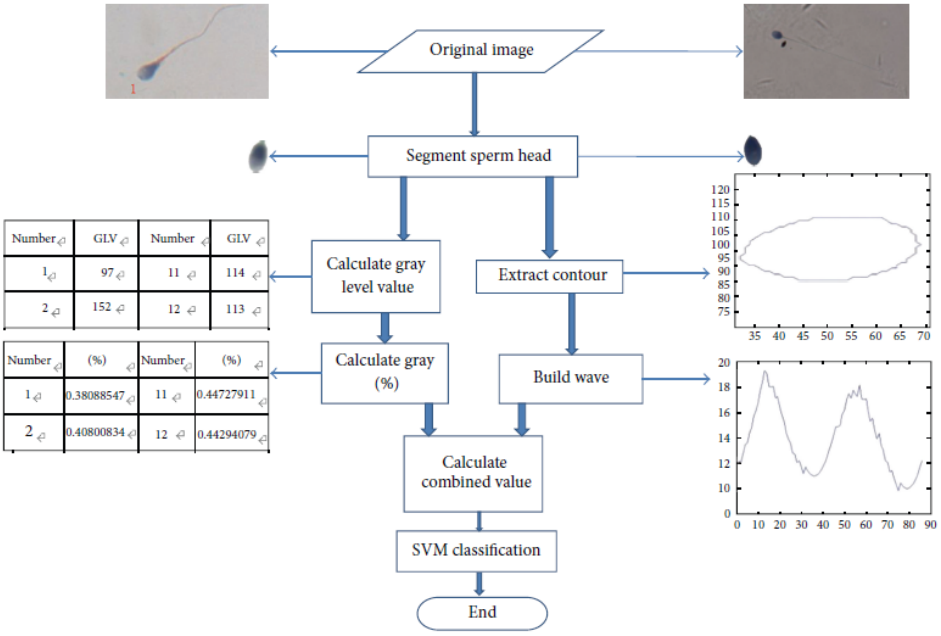

Figura 1.3 Flujo del sistema de clasificación [Tseng, et al., 2013].

En la evaluación realizada, se obtiene un porcentaje de acierto de 87.5% con el método SVM, por lo que se concluye que este acercamiento es factible, y que brinda mejores resultados que los obtenidos por otros métodos existentes [Tseng, et al., 2013].

1.2.6. Conclusiones sobre el estado del arte

Los trabajos presentados en la revisión del estado del arte tienen como objetivo encontrar métodos estandarizados y objetivos que permitan facilitar la tarea de la clasificación de células espermáticas, para disminuir la incertidumbre en el diagnóstico de la infertilidad masculina (Tabla 1.2). Se puede notar el uso de métricas comunes en los distintos estudios para medir la forma de las células, como por ejemplo el alto y ancho de la cabeza de la célula espermática, el porcentaje de acrosoma, la longitud y forma de la cola, el porcentaje de células normales, entre otros. De la misma manera se manifiesta cierta inconformidad con los límites propuestos por la OMS para estas métricas, por considerarlas demasiado estrictas.

Tabla 1.2 Resumen del estado del arte

ESTUDIO	APORTE
Análisis de células espermáticas caninas Con el analizador de Hamilton-Thorne [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]	El analizador de Hamilton-Thorne consiste en un microscopio óptico Olympus, una cámara, un digitalizador de imágenes, una computadora para guardar y analizar los datos recogidos además de un filtro verde para aumentar el contraste entre las células y el fondo. Las pruebas se realizaron sobre células caninas y los resultados obtenidos fueron satisfactorios con relación a los obtenidos con el método manual.
Analizador automático de la morfología de células espermáticas [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007]	El método empleado consta de cinco fases: adquisición de imágenes, detección y extracción de células espermáticas individuales, mejoramiento de la imagen, algoritmo de segmentación, extracción de características y clasificación. Las pruebas se realizaron sobre 216 imágenes microscópicas y los resultados obtenidos fueron satisfactorios.
Analizador automático de la morfología de la cabeza de las células espermáticas usando software libre [Butts, et al., 2011]	Plug-in de código abierto para ImageJ con el que se analizó la morfología de la cabeza de 30 muestras de esperma. Los resultados fueron satisfactorios con relación al método manual.
Máquinas de soporte vectorial aplicadas al diagnóstico de la morfología de las células espermáticas [Tseng, et al., 2013]	Método algorítmico para la clasificación de células espermáticas basado en máquinas de soporte vectorial (algoritmo de clasificación). Se obtuvo un porcentaje de acierto del 87.5%.

Un consenso común es la necesidad de estandarizar los procedimientos y prácticas en todo el proceso de clasificación, desde la manipulación de las muestras, la segmentación, las mediciones, etc. La solución que se propone en el presente trabajo, sigue las recomendaciones propuestas en estudios anteriores, buscando mejorar los resultados obtenidos previamente en cuanto al porcentaje de acierto conseguido. Para ello, se realizaron mejoras en los procesos de extracción de características de las células, así como en la selección del algoritmo de aprendizaje automático y la calibración de sus respectivos parámetros.

1.3. Objetivos y resultados esperados

La presente sección define el objetivo general y los objetivos específicos del presente trabajo de investigación. También se enumeran los resultados esperados por cada objetivo planteado.

1.3.1. Objetivo General

El objetivo general de la presente investigación es desarrollar un modelo computacional que permita caracterizar los aspectos morfológicos de la cabeza de las células espermáticas mediante el procesamiento de micrografías digitales de muestras de esperma y la posterior clasificación de dichas células en normales o anormales mediante técnicas de aprendizaje automático.

1.3.2. Objetivos específicos

A partir del objetivo general, se plantean los siguientes objetivos específicos:

- Objetivo específico 1 - Procesar las imágenes microscópicas de las muestras de células espermáticas para extraer características morfológicas de la cabeza de las mismas.
- Objetivo específico 2 - Seleccionar las mejores características para realizar la clasificación de las células usando métodos automatizados y analíticos.
- Objetivo específico 3 – Clasificar las cabezas de las células espermáticas en normales o anormales de acuerdo a su morfología aplicando un algoritmo de Máquinas de Soporte Vectorial (SVM) a los datos obtenidos en (2) calibrando adecuadamente sus parámetros y evaluando los resultados obtenidos.
- Objetivo específico 4 - Reportar estadísticas sobre la caracterización morfológica de la cabeza de las células y su comparación con los criterios de la OMS.

1.3.3. Resultados esperados

A continuación se exponen los resultados esperados por cada objetivo específico planteado:

- Resultado 1 relacionado al objetivo específico 1: Características de las cabezas de las células espermáticas identificadas en las microimágenes, dadas por mediciones relacionadas a su forma.
- Resultado 2 relacionado al objetivo específico 2: Subconjunto de las características observadas que serán usadas en la clasificación.
- Resultado 3 relacionado al objetivo específico 3: Tasas de acierto y matriz de confusión que muestran el resultado de la clasificación de la cabeza de las células espermáticas en normales o anormales de acuerdo a su morfología.
- Resultado 4 relacionado al objetivo específico 4: Gráficos que muestren la relación entre los resultados obtenidos en la caracterización morfológica de las células espermáticas y los estándares aceptados por la OMS.

1.4. Organización del proyecto

El modelo computacional que se propone en el presente trabajo de investigación tiene como objetivo caracterizar y clasificar la cabeza de las células espermáticas de acuerdo a su morfología, como parte del análisis de evaluación de infertilidad masculina. Dicho modelo consiste en cuatro fases: procesamiento de imágenes micrográficas para la extracción de características, selección de características a partir del análisis de componentes principales (PCA), clasificación morfológica de la cabeza de las células espermáticas en normales o anormales y comparación de la caracterización de las células con relación a los valores de la OMS para validar los resultados obtenidos.

La Figura 1.4 muestra el mapa del documento. En el capítulo 2 se detalla el marco conceptual relacionado al problema biológico del análisis de la fertilidad masculina, dentro del contexto del presente trabajo de investigación. Los capítulos del 3 al 7 muestran a detalle los distintos pasos que componen el modelo computacional que se propone. En el capítulo 8 se discuten los resultados obtenidos y finalmente, el capítulo 9 muestra las observaciones, conclusiones y recomendaciones finales.

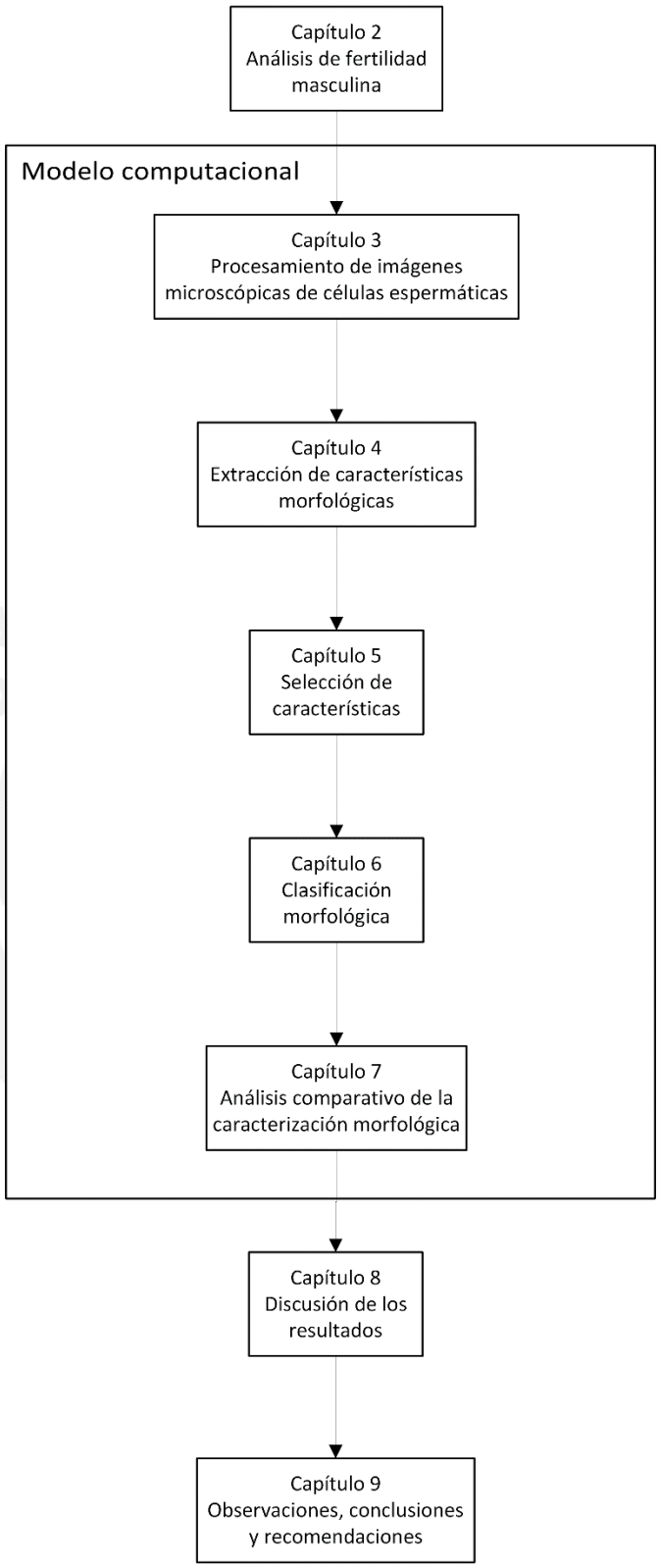

Figura 1.4 Mapa del documento

El capítulo 3 describe las fuentes de datos usadas, en este caso imágenes microscópicas de muestras de células espermáticas. El capítulo 4 detalla el primer paso del modelo que es la extracción de las características morfológicas de la cabeza de las células espermáticas de acuerdo a lo planteado en el resultado esperado número 1. En el capítulo 0 se aborda la selección de las características que mejor identifiquen a las células analizadas a través de la aplicación de análisis de componentes principales (PCA). Esto corresponde al resultado esperado número 2.

Una vez se cuente con el conjunto de datos que caracteriza a la cabeza de las células espermáticas se puede proceder a la clasificación de las mismas en normales o anormales empleando técnicas de aprendizaje de máquina y evaluando los resultados adecuadamente. Este es el contenido del capítulo 6, relacionado al resultado esperado número 3.

El último paso del modelo se detalla en el capítulo 7 y consiste en un reporte comparativo probabilístico de las características observadas en las células espermáticas normales con relación a los rangos que establece la OMS. Este capítulo corresponde al resultado esperado número 4.

Como herramienta computacional se empleó el software Mathematica de Wolfram [Wolfram, 2014] por proveer un entorno propicio para la experimentación y el procesamiento de imágenes y datos. Además, Mathematica cuenta con un amplio conjunto de algoritmos ya implementados en el área del procesamiento de imágenes, aprendizaje automático, análisis estadístico entre otros.

2. ANÁLISIS DE FERTILIDAD MASCULINA

El objetivo del presente capítulo es el entendimiento del problema del análisis de fertilidad masculina desde el punto de vista biológico, haciendo énfasis en el análisis morfológico de las células espermáticas. Para ello, se exponen los componentes de las células espermáticas, el proceso de análisis de fertilidad y análisis morfológico así como los criterios de normalidad y anormalidad que se tienen en cuenta a la hora de clasificar células espermáticas.

2.1. Espermiograma

En el caso de los hombres, existen diferentes métodos para el análisis de células espermáticas, siendo el más usado, el espermiograma. De acuerdo a la OMS, el espermiograma consta de dos fases fundamentales: un examen macroscópico de las muestras, donde son analizadas algunas características físicas como la viscosidad, el olor, el pH y el aspecto. El otro examen es el microscópico, donde se miden diversos parámetros de los espermatozoides, como concentración, movilidad y morfología [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007]. De todos estos parámetros, el análisis morfológico ha mostrado ser el más acertado en cuanto a la predicción del potencial de fertilización de un hombre [Nikolettos, et al., 1999].

2.2. Caracterización de células espermáticas

La célula espermática está compuesta por una cabeza, cuello, pieza media y cola (Figura 2.1). Una membrana de plasma cubre la cabeza y llega hasta la punta de la cola. La mayor parte de la cabeza está compuesta por un núcleo compacto, que contiene el ADN, y que a su vez es cubierto por un acrosoma y la membrana de plasma. El acrosoma cubre más de las dos terceras partes del núcleo de la célula espermática y su función es que el espermatozoide pueda penetrar en el óvulo y se efectúe la fecundación [Nikolettos, et al., 1999].

La pieza media es la fuerza impulsora del espermatozoide. Las fibras que la componen facilitan el movimiento de la esperma a la vez que le brindan protección en su tránsito por los tractos reproductivos del hombre y la mujer [Nikolettos, et al., 1999]. La cola tiene la parte del principio, y la parte final y está cubierta por una densa fibra que reduce su grosor mientras se acerca al final de la cola [Nikolettos, et al., 1999].

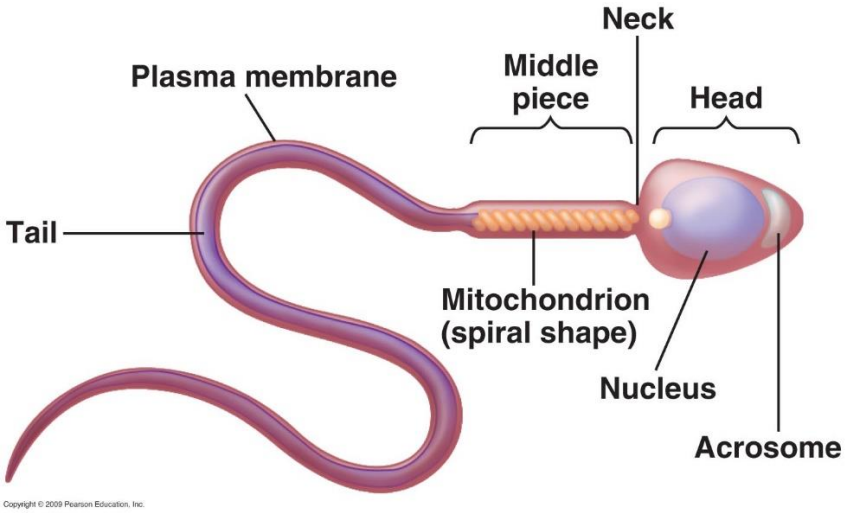

Figura 2.1 Estructura de una célula espermática [Pearson Education Inc., 2009].

2.3. Clasificación de células espermáticas

Definir la morfología de células espermáticas normales es una tarea muy difícil por la gran heterogeneidad de las mismas (Figura 2.2). La OMS emplea para este fin las relaciones entre el porcentaje de formas normales y algunas tasas de fertilidad, como el tiempo para embarazo, el éxito en fertilizaciones in vivo e in vitro, etc. Según esta idea, el rango de porcentaje de células normales tanto para hombres fértiles como infértiles está entre 0 y 30% [World Health Organization, 2010].

Según la OMS, los siguientes son los requisitos que debe cumplir una célula espermática para ser considerada normal. Todas las formas limítrofes deben ser consideradas como anormales [World Health Organization, 2010]:

- La cabeza debe tener forma regular y ovalada. Debe haber una región acrosomal bien definida, que comprenda del 40 al 70% del área de la cabeza.
- La pieza media debe ser fina, regular y debe tener aproximadamente la misma longitud que la cabeza.
- La cola debe ser uniforme y más delgada que la pieza media, con una longitud aproximada de 45µm (aproximadamente 10 veces la longitud de la cabeza).

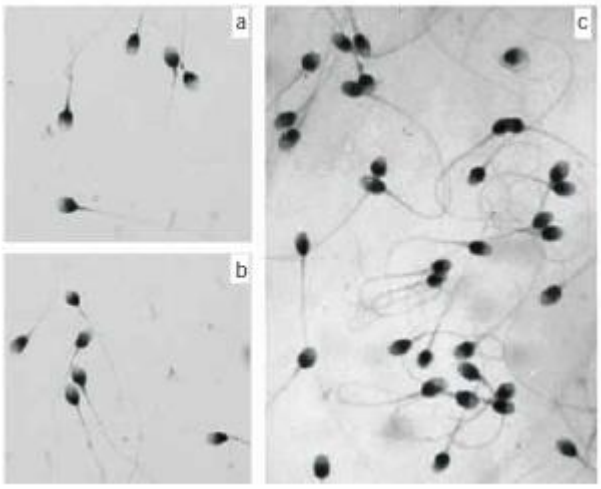

Figura 2.2 Células normales [World Health Organization, 2010].

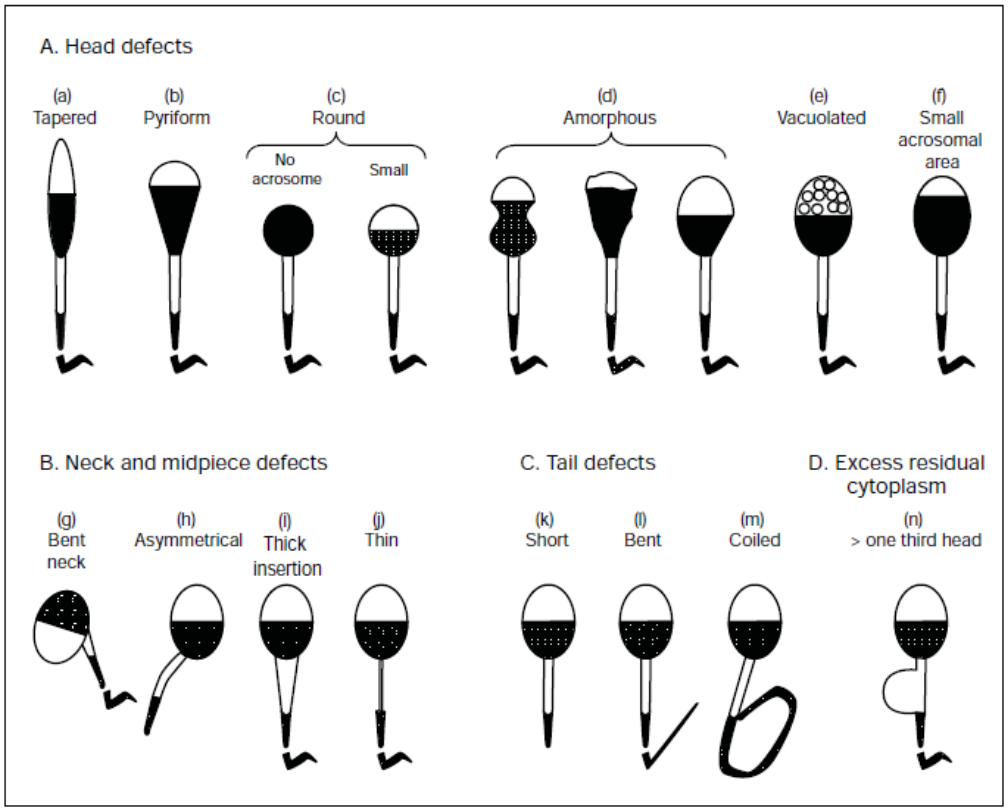

Figura 2.3 Dibujos esquemáticos de algunas de las anomalías que pueden presentar las células espermáticas [World Health Organization, 2010].

Las categorías de anomalías que define la OMS [World Health Organization, 2010], se describen a continuación. Algunas de estas anomalías se muestran en la Figura 2.3.

- Defectos en la cabeza: Cabeza muy pequeña o muy grande, con forma piriforme, redonda, amorfa, doble cabeza, o cualquier combinación de los anteriores.
- Defectos en el cuello y pieza media: Inserción asimétrica de la pieza media en la cabeza. Pieza media gruesa o irregular, doblada, muy delgada, o una combinación de los anteriores.
- Defectos en la cola: Corta, múltiple, rota, enroscada, con ángulos puntiagudos, grosor irregular, o cualquier combinación de las anteriores.
- Exceso de citoplasma residual (ERC): Asociado a un proceso anormal de formación de esperma. Se caracteriza por grandes cantidades irregulares de citoplasma, un tercio o más del tamaño de la cabeza.

3. PROCESAMIENTO DE IMÁGENES MICROSCÓPICAS DE CÉLULAS ESPERMÁTICAS

En el presente capítulo se describen las fuentes de datos usadas en la investigación, en este caso imágenes microscópicas de células espermáticas. Asimismo se explica el algoritmo de Otsu aplicado para procesar las imágenes de modo que fuese posible la extracción de características.

3.1. Fuentes de microimágenes

La primera fuente de imágenes fue un repositorio de micrografías digitales puesto a disposición en el Centro de Espermiogramas Digitales Asistidos por Internet (CEDAI) de Chile [Chang, et al., 2014]. Este repositorio forma parte de una investigación realizada por el Departamento de Ciencias de la Computación de la Universidad de Chile en colaboración con otras instituciones de este país [Chang, et al., 2014]. Dicha investigación propone un estándar para la segmentación de la cabeza de la célula espermática y las imágenes que se obtuvieron son resultado de aplicar el método por ellos propuesto.

En este caso, para cada micrografía digital hay asociada una imagen binaria (solo color blanco y negro), que constituye una máscara de la imagen original, donde aparecen resaltadas en color blanco las cabezas de las células espermáticas. La imagen original se observa en la Figura 3.1 y su respectiva máscara en la Figura 3.2.

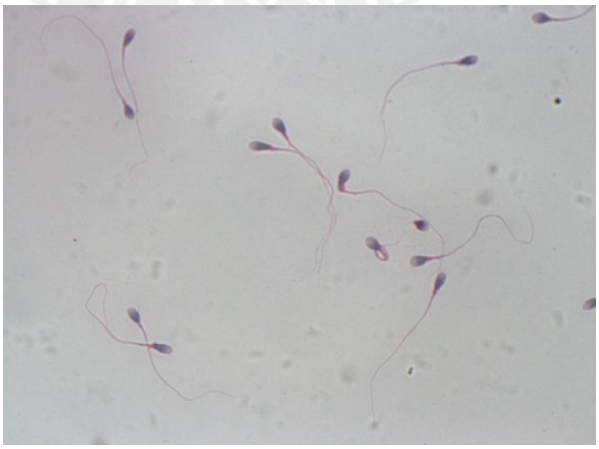

Figura 3.1 Ejemplo de imagen de la primera fuente [Chang, et al., 2014].

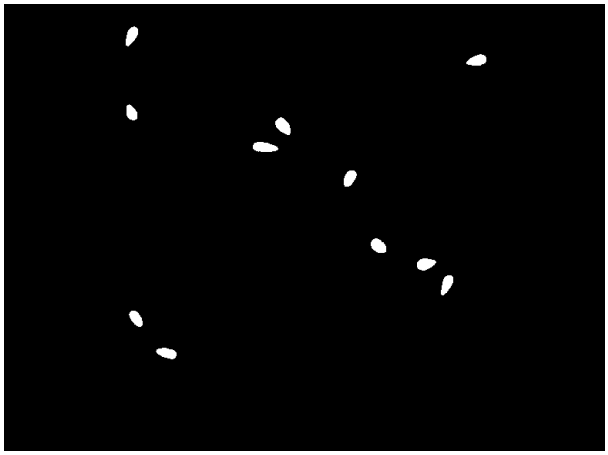

Figura 3.2 Ejemplo de máscara de la cabeza de las células espermáticas [Chang, et al., 2014].

De esta fuente se obtuvieron 19 imágenes con una resolución de 780 por 580 pixeles, que contienen un total de 240 células espermáticas, de las cuales 34 eran normales y el resto anormales. Para las células anormales, la fuente incluía la especificaciín del tipo de anomalía. La Tabla 3.1 muestra los totales de células en cada categoría. Dicha clasificación fue realizada por los especialistas en espermiogramas que participaron en la investigación desarrollada por la Universidad de Chile [Chang, et al., 2014].

Tabla 3.1 Total de células por categoría.

Etiqueta	Total	Descripción
0	34	Normal
1	81	Cónica
2	50	Piriforme
3	0	Sin acrosoma
4	14	Microcefálica
5	1	Macrocefálica
6	0	Bicefálica
7	68	Amorfa
8	6	Vacoulada
9	0	Área acrosomal grande
10	1	Área acrosomal pequeña
11	2	Otros

La cantidad de imágenes obtenidas de esta primera fuente no era suficiente, debido a la poca cantidad de células normales identificadas y a la dificultad de los algoritmos de aprendizaje de máquina de construir modelos predictivos eficaces cuando no existe una

muestra uniforme de las clases que se quieren representar (en este caso cabezas normales o anormales). Por esta razón, se tuvo que recurrir a una segunda fuente de imágenes.

El manual de la OMS, en su quinta edición [World Health Organization, 2010] incluye una serie de ejemplos de imágenes de muestras de células espermáticas y su clasificación en normales o anormales de acuerdo a la morfología de la cabeza, pieza media y cola. Un ejemplo de esto puede observarse en la Figura 3.3. Dicha clasificación, según se enuncia en el propio manual, fue realizada por el experto Dr. Thinus Kruger [World Health Organization, 2010]. Del manual fueron extraídas 28 imágenes que contienen un total de 226 células espermáticas de las cuales 128 eran normales y 98 anormales.

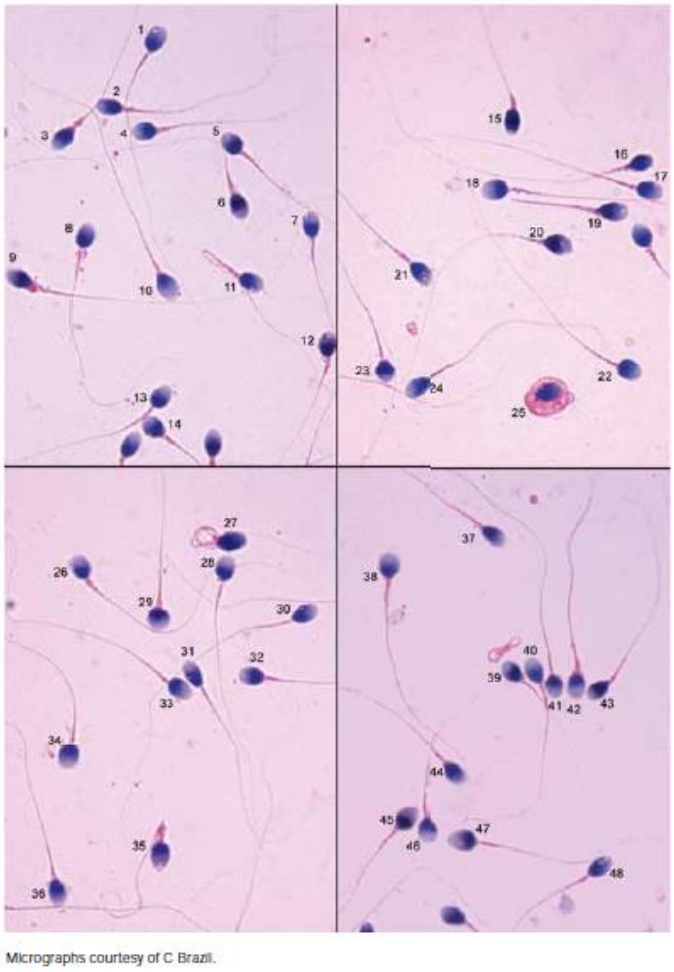

Figura 3.3 Imagen extraída de la segunda fuente [World Health Organization, 2010].

3.2. Procesamiento de imágenes

Para las imágenes extraídas de [Chang, et al., 2014], no fue necesario realizar ningún procesamiento preliminar, no ocurriendo así para las imágenes extraídas del manual de la OMS [World Health Organization, 2010]. En este caso fue necesario procesar las imágenes a fin de identificar la cabeza de las células espermáticas de manera similar a como se tenía en la primera fuente de datos. Esta tarea se realizó con la ayuda de la herramienta Mathematica de Wolfram [Wolfram, 2014] y los métodos empleados se detallan a continuación.

3.2.1. Método de Otsu

El método de Otsu [Otsu, 1979] se utiliza en visión computacional y procesamiento de imágenes para reducir una imagen en escala de grises a una imagen binaria. Generalmente se usa cuando se quiere separar un objeto de su fondo, como quiere hacerse en este caso con las cabezas de las células espermáticas. Para ello, se asume que la imagen tiene dos tipos de pixeles, los del fondo y los del frente (o del objeto que se quiere resaltar). El algoritmo calcula un umbral óptimo que separe estos dos tipos de pixeles de modo que se minimice la varianza dentro de una misma clase (tipo de pixel de fondo o frente). La Figura 3.4 muestra un ejemplo de la aplicación del método de Otsu a una imagen en escala de grises.

Figura 3.4 Ejemplo de funcionamiento de método de Otsu sobre imagen en escala de grises [Wikipedia, 2014].

3.2.2. Binarización de las imágenes usando Mathematica

Mathematica contiene diversas funciones para realizar procesamiento de imágenes, una de ellas es Binarize [Wolfram, 2014]. Esta función está preparada para procesar todo tipo de imágenes, convirtiéndolas primero a escala de grises y luego aplicando el método de Otsu, descrito en la sección anterior.

La función Binarize se aplicó a las imágenes extraídas del manual de la OMS. La Figura 3.5 muestra un ejemplo del resultado obtenido. En este ejemplo se observa que en algunos casos, la binarización no solo resalta en blanco las cabezas de las células espermáticas sino parte de la pieza media. Como el objetivo era realizar una caracterización morfológica solamente de la cabeza de la célula espermática, hubo que realizar ajustes manuales utilizando un editor de imágenes, en este caso del GNU Image Processing Program [GIMP, 2014]. La Figura 3.6 muestra cómo queda la imagen luego de realizar el tratamiento manual.

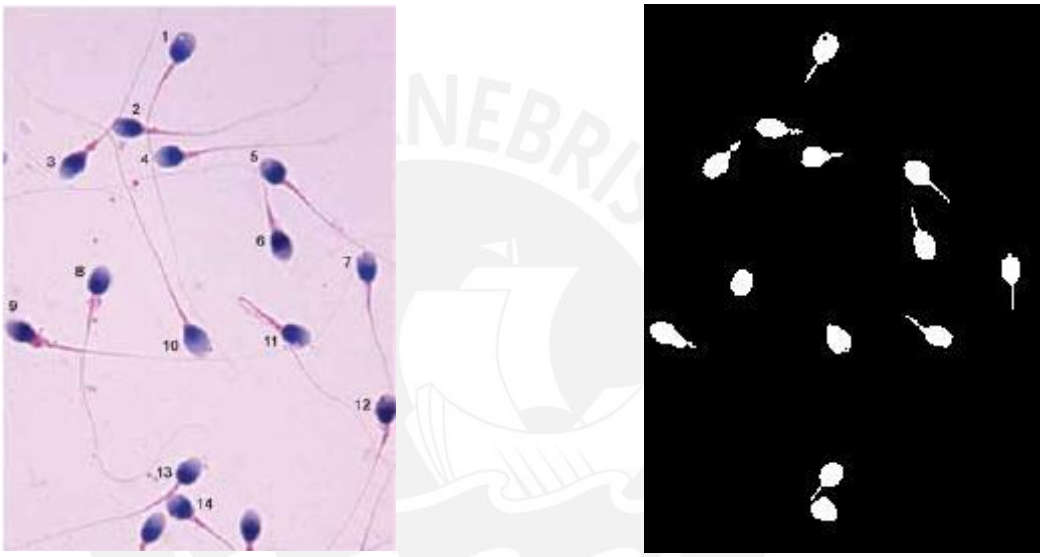

Figura 3.5 Resultado de la aplicación de la función Binarize de Mathematica.

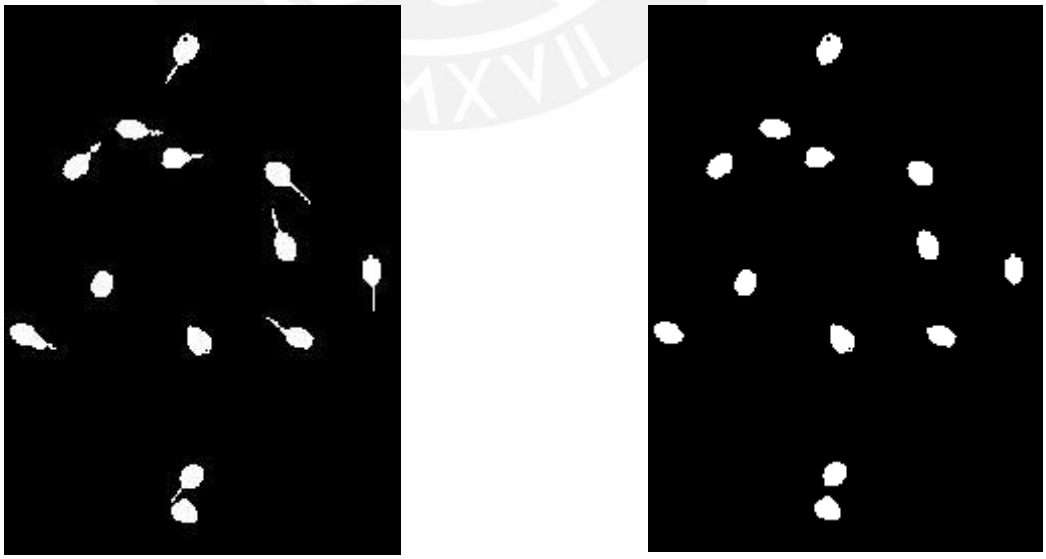

Figura 3.6 Resultado de la corrección manual de la imagen.

Este ajuste fue realizado a todas las imágenes procesadas por Binarize. Del total de 226 células presentes originalmente, se pudieron identificar las cabezas de 221 de ellas, o sea un 97.8%.

4. EXTRACCIÓN DE CARACTERÍSTICAS MORFOLÓGICAS

En el presente capítulo se aborda el tema de la caracterización de la cabeza de las células espermáticas. Para ello se enumeran y justifican las métricas tomadas en cuenta, los métodos y herramientas utilizados.

4.1. Métricas

La extracción de las características de la cabeza de las células espermáticas es el paso más importante del modelo computacional que se propone en la presente investigación. Los errores cometidos en esta etapa, ya sea por la selección de métricas no pertinentes o por el registro de valores incorrectos de dichas métricas, va a traer como consecuencia que dichos errores se propaguen a las restantes fases del proceso y por ende se reflejen en el resultado final obtenido, restándole validez a la investigación.

No existe un consenso en las investigaciones revisadas en cuanto a cuáles son las características morfológicas que deben tomarse de la cabeza de las células espermáticas. Sin embargo, hay algunas que se repiten en diversos estudios como el largo de la cabeza, el ancho, el área y el perímetro [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007; Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]. Por esta razón se incluyeron estas cuatro características dentro del conjunto de métricas empleadas.

Según el manual de la OMS [World Health Organization, 2010], una cabeza normal debe tener forma ovalada y contornos regulares. Para determinar la forma de la cabeza no es suficiente medir el largo, ancho, área y perímetro de la misma; sino que existen otras métricas que pueden ser tomadas en cuenta, como la excentricidad de la elipse que se aproxima a la forma de la cabeza de la célula, el ratio entre el ancho y el largo, la simetría, etc. Por otro lado, tener en cuenta la regularidad de los contornos puede ser engañoso, debido a que la suavidad de los bordes puede verse comprometida frente a diversos factores como la resolución de la imagen, el método que se haya usado para la detección de las cabezas de las células espermáticas, entre otros. Por las razones antes expuestas, se consideró incluir la excentricidad y el ratio entre ancho y largo dentro de las métricas para caracterizar la cabeza de las células espermáticas.

En la Tabla 4.1 se listan las características morfológicas que fueron extraídas de las microimágenes con el objetivo de caracterizar las cabezas de las células espermáticas que en ellas aparecen. Nótese que dichas características fueron definidas de manera experimental, siguiendo las recomendaciones de la OMS [World Health Organization, 2010] y los resultados obtenidos en estudios consultados en el estado del arte [Rijsselaere, Van Soom, Hoflack, Maes, & de Kruif, 2004]. Los valores para las características medidas fueron obtenidos de manera automática con la ayuda de la herramienta Mathematica.

Tabla 4.1 Características a extraer.

Propiedad	Unidad de Medida
Largo	µm
Ancho	µm
Ratio (Ancho:Largo)	–
Área	μm^2
Perímetro	µm
Excentricidad	–

4.2. Detección automática de componentes usando Mathematica

En Mathematica es posible realizar la detección automática de los componentes de una imagen en 2D utilizando la función ComponentMeasurements [Wolfram, 2014]. En el caso de una imagen binarizada, esta función identifica los componentes de la imagen basado en los pixeles con valor distinto de cero (blancos) que se encuentren conectados (adyacentes), como se muestra en la Figura 4.1. Para cada uno de los componentes identificados, Mathematica puede calcular una serie de propiedades relacionadas al área, al perímetro, a la elipse que mejor se ajuste a la forma, entre otros.

Para la primera fuente de datos, Mathematica identificó 201 de las 240 células existentes, o sea un 83.75%, de las cuales 29 eran normales y 172 anormales. Para la segunda fuente de datos, se identificaron todas las 221 células detectadas en el proceso de binarización descrito en el capítulo anterior. De ellas, 126 eran normales y 95 anormales.

Además de las propiedades obtenidas de manera automática, se calculó el ratio de ancho entre largo de la cabeza de la célula. Con esta característica se completó la caracterización morfológica de la cabeza de las células espermáticas.

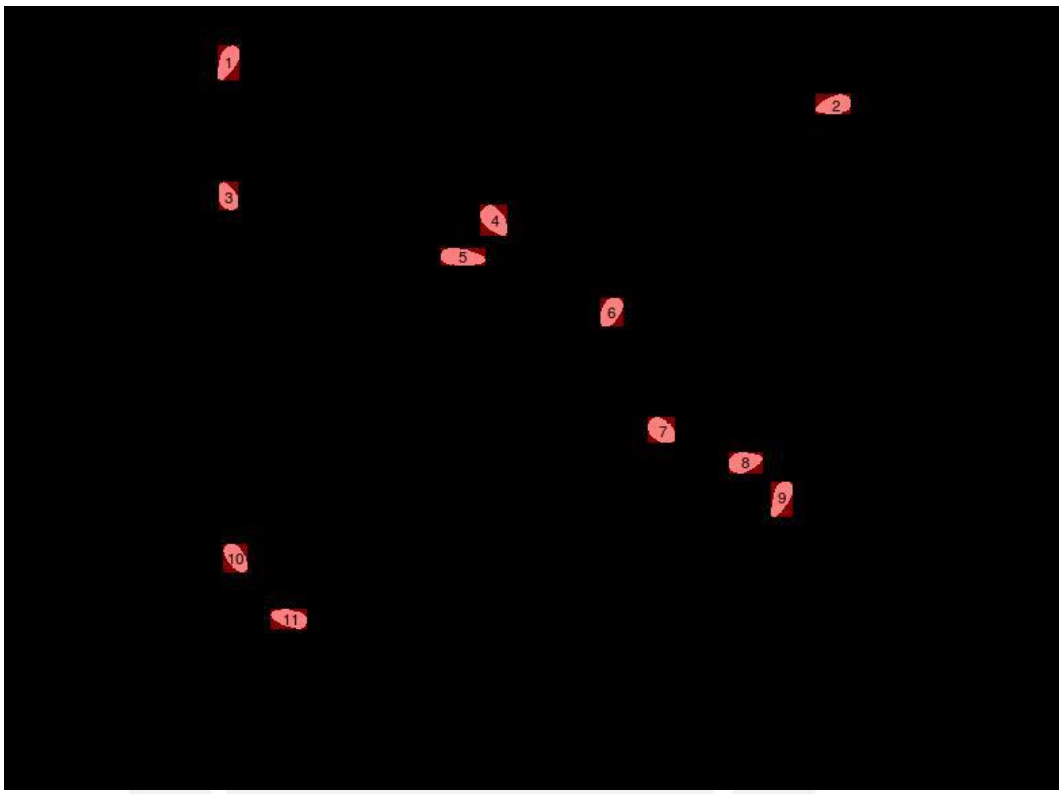

Figura 4.1 Identificación de componentes en Mathematica.

4.3. Estandarización de datos

Dado que las mediciones realizadas estaban en pixeles, se procedió a estandarizar las mismas, convirtiendo a micrómetros teniendo en cuenta las dimensiones y la resolución de las imágenes. Para el primer grupo de imágenes el factor de conversión fue de 0.21 micrómetros por pixel, y para el segundo grupo de imágenes es de 0.28 micrómetros por pixel.

Adicionalmente, fue necesario añadir a estos datos los valores de las etiquetas, o sea, la clasificación otorgada por los especialistas y contenida en cada una de las fuentes de datos. La Tabla 4.2 muestra la distribución de células en cada una de las categorías para la primera fuente de datos [Chang, et al., 2014]. Nótese que la distribución no es uniforme y que existen algunas categorías con menos de diez representaciones en el conjunto de datos. Además el objetivo de la clasificación es distinguir entre células normales y anormales, por lo que se realizó un mapeo entre estas categorías, donde se unificaron

todas las anomalías bajo una sola categoría: anormal, quedando la distribución como se muestra en la Tabla 4.3. En el caso de la segunda fuente de datos [World Health Organization, 2010] no fue necesario realizar este mapeo pues solo se consideraron dos alternativas para la cabeza de las células: normal o anormal.

Tabla 4.2 Distribución de células por categoría para la primera fuente de datos.

sin clasificación	9
normal	29
cónica	68
piriforme	38
microcefálica	9
macrocefálica	1
amorfa	51
problemas de vacuolas	5

Tabla 4.3 Distribución final de células por categoría para la primera fuente de datos después del mapeo de etiquetas.

Total	201
Normales	29
Anormales	172

Tabla 4.4 Ejemplo de datos extraídos para la primera fuente de datos.

Largo	Ancho	Ratio	Área	Perímetro	Excentricidad	Clase
5.08747	2.73572	0.248262	11.0195	1.06837	0.843112	1
5.81677	2.71025	0.219	12.3756	1.22152	0.884818	1
5.90783	2.60822	0.21804	11.9621	1.24064	0.897268	1
4.60951	3.07419	0.279537	10.9974	0.967997	0.745127	0
5.91545	1.8453	0.213624	8.63809	1.24224	0.9501	1
4.4346	2.78366	0.284973	9.76815	0.931265	0.778444	0
4.57568	2.55032	0.276535	9.22241	0.960892	0.830268	0
5.43746	3.02513	0.233621	12.9489	1.14187	0.830948	1
5.46545	3.1634	0.239208	13.2245	1.14774	0.815469	1
7.1534	2.30281	0.18147	12.6898	1.50221	0.946767	1
5.45348	2.72765	0.232853	11.7141	1.14523	0.865929	1

La Tabla 4.4 muestra un ejemplo de 10 registros, cada uno correspondiente a una célula identificada en las muestras microscópicas. Los valores que se muestran se encuentran expresados en micrómetros.

5. SELECCIÓN DE CARACTERÍSTICAS

Sobre la selección inicial de características extraídas de las microimágenes, se realizó Análisis de Componentes Principales (PCA). En términos generales, este método permite extraer información relevante de conjuntos de datos que pueden parecer confusos, lo cual permite reducir la dimensionalidad del problema, o sea la cantidad de variables; así como revelar estructuras que pueden encontrarse escondidas dentro de los datos, como por ejemplo, relaciones no triviales entre los mismos, datos redundantes, etc. [Shlens, 2003].

A continuación se describe el método PCA y su aplicación al conjunto de datos obtenido hasta el momento con la ayuda de la herramienta Mathematica.

5.1. Análisis de componentes principales

Un componente principal puede ser definido como una combinación lineal de ciertas variables observadas [Hatcher, 1994]. A continuación se expone un método a partir de un ejemplo donde se tienen dos variables observadas x e y. La Figura 5.1 muestra los datos originales y el gráfico correspondiente. Este ejemplo ha sido adaptado de [Smith, 2002].

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

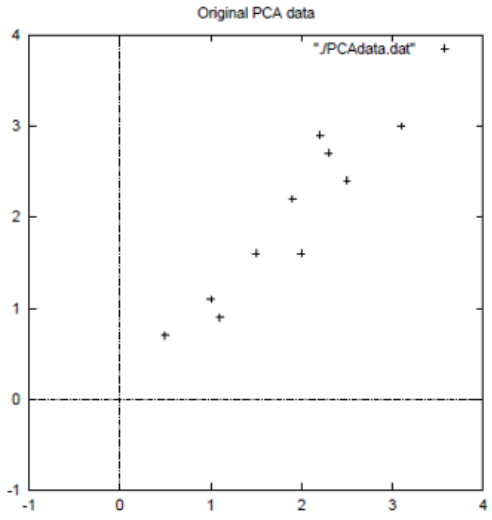

Figura 5.1 Datos originales [Smith, 2002].

El primer paso es centrar los datos, esto quiere decir restar las medias. Para ello es necesario calcular la media de cada variable, y luego sustraer todos los valores. Esto produce un nuevo conjunto de datos, con media cero, que se observa en la Figura 5.2.

$$\bar{x} = 1.81 \quad \bar{y} = 1.91$$

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Figura 5.2 Datos centrados [Smith, 2002].

El siguiente paso es calcular la matriz de covarianza:

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

El tercer paso es calcular los vectores y valores propios de la matriz de covarianza:

$$valores\ propios = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$vectores\ propios = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Este proceso permite obtener líneas que caractericen los datos, como se muestra en la Figura 5.3.

Ahora, se deben seleccionar los componentes y formar el nuevo vector de características. Se puede observar que los valores propios son diferentes. El vector propio asociado al valor propio de mayor valor es el componente principal del conjunto de datos, al ser el que mejor refleja la relación entre las dos variables x e y. Esto se puede visualizar en la Figura 5.3.

En general, el procedimiento consiste en ordenar los valores propios de mayor a menor para obtener los componentes ordenados por su nivel de significancia. Aquí es donde se seleccionan tantos componentes como sea necesario. Siempre se perderá información, pero si los valores propios asociados a estos componentes que no se están tomando son pequeños, entonces esta información no es tan significativa [Smith, 2002].

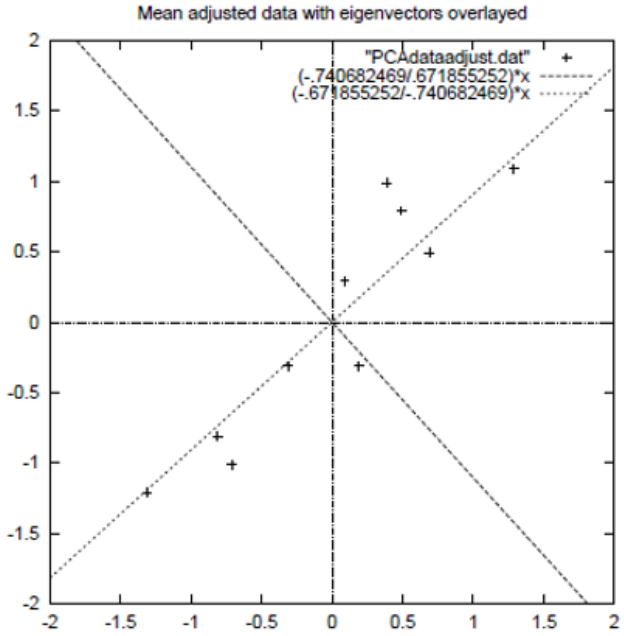

Figura 5.3 Datos centrados y vectores propios [Smith, 2002].

Si inicialmente se tuvieran n variables, se obtendrían n valores y vectores propios. Si se seleccionan los primeros p (ordenados por nivel de significancia), entonces el conjunto de datos finales va a tener p dimensiones [Smith, 2002].

Con los vectores seleccionados se forma el nuevo vector de características. En el caso de este ejemplo, si seleccionamos solo una de las características:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

El paso final es derivar el nuevo conjunto de datos a partir del vector de características formado. Para ello se multiplica el vector de características transpuesto por el conjunto de datos originales, transpuesto. Esto es:

$$Datos\ finales = VectorCaracterísticasFila \times DatosAjustadosFila$$

Donde $VectorCaracterísticasFila$ es el vector de características transpuesto, donde ahora los vectores propios son filas, y la fila más arriba es la que representa al vector más significativo, y $DatosAjustadosFila$ es la matriz que contiene los datos iniciales centrados, transpuesta. Esto da como resultado la data inicial expresada solamente en términos de los vectores seleccionados. Los datos iniciales estaban expresados en relación a las variables x e y, ahora están expresados en términos del vector de características seleccionado [Smith, 2002].

5.2. Unificación y estandarización de los datos

Antes de aplicar PCA al conjunto de datos obtenido en el capítulo 4 se procedió a unificar las dos fuentes de datos en una sola, a eliminar las instancias no clasificadas y a estandarizar los datos. La Tabla 5.1 muestra la distribución de células normales y anormales en el conjunto de datos unificado. Las instancias no clasificadas que fueron eliminadas (en total 49) eran instancias que no habían sido clasificadas en las fuentes de datos originales.

Tabla 5.1 Distribución de células normales y anormales en el conjunto de datos unificado.

Total	422
Normales	155
Anormales	267

En cuanto a la estandarización de los datos, se utilizó la función Standarize de Mathematica [Wolfram, 2014] que modifica y escala los datos de modo que tengan media 0 y varianza 1. Este procedimiento se recomienda para obtener mejores resultados en la clasificación utilizando el algoritmo SVM [Chih-Wei, Chih-Chung, & Chih-Jen, 2010]. Una de las ventajas de estandarizar los datos es otorgar igualdad de oportunidades a todas las características, evitando que aquellas de mayor valor numérico dominen a las de menor valor numérico. Además, la estandarización reduce la pérdida de información en los cálculos numéricos [Chih-Wei, Chih-Chung, & Chih-Jen, 2010].

5.3. Análisis de componentes principales utilizando Mathematica

Mathematica cuenta con la función PrincipalComponents [Wolfram, 2014] la cual transforma una matriz de datos de entrada en su matriz de componentes principales, transformando las columnas de la matriz original en columnas no relacionadas, ordenadas de izquierda a derecha en orden decreciente de varianza. En principio, la matriz que devuelve Mathematica tiene la misma dimensión que la matriz que se suministra como entrada a la función. Si se quisieran obtener las n componentes principales, bastaría con seleccionar las n primeras columnas de la matriz resultante.

Con el objetivo de determinar cuántos componentes se iban a seleccionar, se realizó un análisis comparativo, probando la efectividad del clasificador SVM con el conjunto de

datos resultante de tomar la primera componente principal, luego las dos primeras, luego 3 y así sucesivamente. La Figura 5.4 muestra los resultados obtenidos.

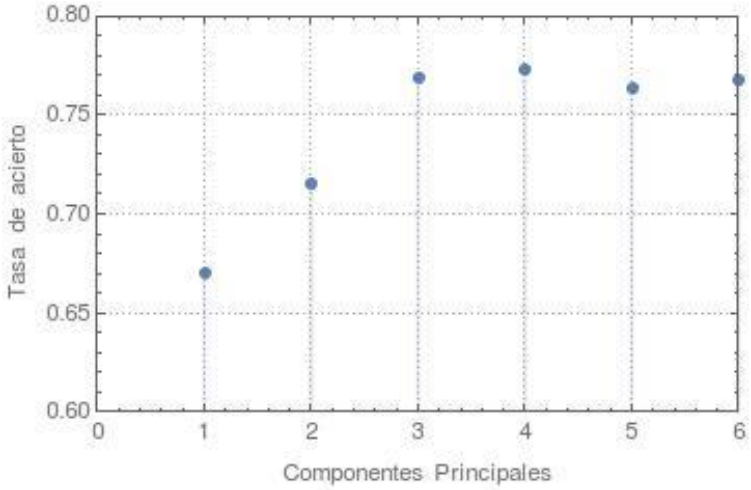

Figura 5.4 Tasa de acierto en función de la cantidad de componentes.

En el gráfico se observa que a partir de 3 componentes, el resultado de la clasificación se estabiliza alrededor del 77%. Esto ocurre ya que los restantes componentes no ofrecen información relevante en cuando a la diferenciación entre células normales y anormales. De esta manera, el conjunto de datos con el que se trabajó, contiene las 3 componentes más importantes, según el análisis PCA realizado.

6. CLASIFICACIÓN MORFOLÓGICA

En el presente capítulo se abordan los métodos y procedimientos relativos a la clasificación morfológica de las células espermáticas: aprendizaje de máquina y algoritmo de clasificación de máquinas de soporte vectorial (SVM). Asimismo, se muestra la aplicación de dichos métodos al conjunto de datos obtenido en el capítulo anterior.

6.1. Métodos utilizados en la clasificación

A pesar de que la caracterización de las células espermáticas solo ha comprendido las características morfológicas de la cabeza, realizar una clasificación preliminar en normal o anormal puede ser de gran utilidad. En primer lugar una célula cuya cabeza no presente una morfología normal, no podrá ser clasificada como normal bajo ningún otro concepto [World Health Organization, 2010], y podrá ser descartada. Por otro lado, si la célula es clasificada como normal, debe entenderse que se refiere a que la morfología de su cabeza es normal, sin tener en cuenta el contenido de acrosoma, y que esta célula debe seguir siendo analizada para poder llegar a un diagnóstico final.

En la presente investigación se ha empleado un enfoque de aprendizaje de máquina estadístico, donde cada entidad (célula espermática) ha sido representada con un conjunto de características que la definen. Dentro de los algoritmos que siguen este enfoque, se seleccionó el de Máquina de Soporte Vectorial (SVM) por ser ampliamente usado en aplicaciones bioinformáticas donde se procesan imágenes obteniéndose buenos resultados [Chapelle, 1998].

6.1.1. Aprendizaje de máquina

El aprendizaje de máquina es una rama de la inteligencia artificial, que engloba una serie de procedimientos y algoritmos que son capaces de “aprender” de sus datos de entrada. El siguiente esquema muestra el diagrama de un sistema de aprendizaje básico (Figura 6.1). En sentido general, el algoritmo de aprendizaje de máquina recibe como entrada un conjunto de datos de entrenamiento, que generalmente son resultado de observaciones del fenómeno que se quiere estudiar o predecir. La salida del algoritmo es un modelo predictivo, al que se le suministran nuevos datos de entrada, esperándose obtener una predicción sobre el comportamiento o significado de dichos datos [van Leeuwen, 2004].

Existen un conjunto de conceptos asociados a los modelos de aprendizaje [van Leeuwen, 2004]:

- Aprendiz: Quién o qué va a realizar el aprendizaje. Los algoritmos de aprendizaje de máquina, pueden incluirse dentro de sistemas más generales.

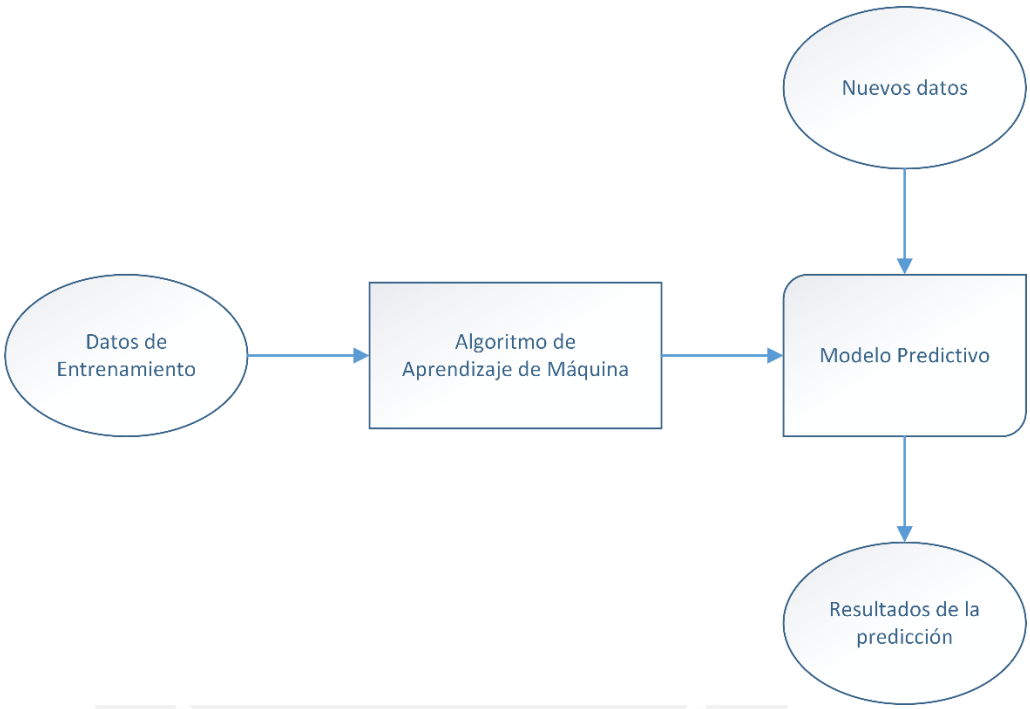

Figura 6.1 Aprendizaje de Máquina.

- Dominio: Qué, o sobre qué se quiere aprender. Existen muchas posibilidades: conceptos, juegos, idiomas, etc.
- Objetivo: Por qué se realiza el aprendizaje. Puede realizarse para extraer una serie de reglas de un conjunto de datos dispersos, para ejercer control sobre un sistema, entre otros.
- Representación: La forma en que se van a representar los objetos que quieren ser aprendidos.
- Tecnología algorítmica: El marco de trabajo a utilizar: redes neuronales, árboles de decisión, redes probabilísticas, aprendizaje basado en reglas, máquinas de soporte vectorial, etc. Pueden utilizarse estrategias multiobjetivo.
- Fuente de información: La información que va a ser usada para entrenar. Pueden ser respuestas a preguntas, resultados de determinadas

observaciones de un fenómeno dado, etc. Se dice que una fuente de información tiene ruido cuando tiene errores.

- Escenario de entrenamiento: Esta es la descripción del proceso de aprendizaje. Pueden haber escenarios interactivos (on-line) y otros en que los datos de ejemplos se proveen una vez solamente (off-line). También se realiza una distinción entre aprendizaje supervisado y aprendizaje no supervisado. El primero se refiere a la predicción de respuestas que ya se conocen de antemano, mientras que en el segundo, el programa debe determinar ciertas regularidades en los datos para predecir la respuesta por él mismo.
- Conocimiento previo: Lo que se sabe a priori, pueden ser propiedades matemáticas, o conceptos relativos a la naturaleza del problema que se quiera resolver: conceptos biológicos, médicos, etc. Esto puede facilitar el proceso de aprendizaje, y hacer que el programa converja más rápidamente a una respuesta.
- Criterio de éxito: Criterio que muestre si el aprendizaje fue exitoso o no. Dependiendo del tipo de problema, se va a requerir un resultado más o menos preciso. Las tasas de éxito se miden a partir de pruebas realizadas al modelo con datos de prueba.
- Performance: Recursos computacionales en términos de tiempo, espacio, y capacidad de procesamiento requerida para determinada tarea. También se puede medir por la precisión alcanzada en el proceso, o sea, qué tan bien ha aprendido esta máquina.

6.1.2. Máquinas de soporte vectorial

Las máquinas de soporte vectorial son un modelo de aprendizaje supervisado usado para el reconocimiento de patrones y la clasificación. Tiene numerosas aplicaciones en el campo de la Bioinformática, debido a sus altos niveles de precisión, su habilidad de manejar data multidimensional y su flexibilidad [Ben-Hur & Weston, 2010].

Este clasificador resuelve el problema de clasificación para dos clases usando un modelo linear de la forma:

$$y(x) = w^T \phi(x) + b$$

La función $\phi(x)$ representa una transformación del espacio de características, w es un vector de parámetros y b es un parámetro de parcialidad. Los datos de entrenamiento

están formados por N vectores $x_1, x_2, \dots x_N$, que corresponden a los valores $t_1, t_2, \dots t_N$, donde $t_n \in \{-1,1\}$ y los nuevos datos x serán clasificados de acuerdo al signo de la función $y(x)$ [Bishop, 2006] (Figura 6.2).

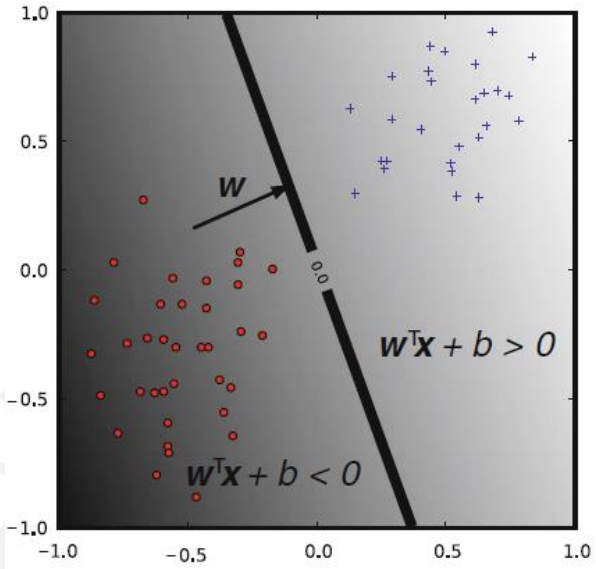

Figura 6.2 Representación del clasificador SVM con núcleo lineal [Ben-Hur & Weston, 2010].

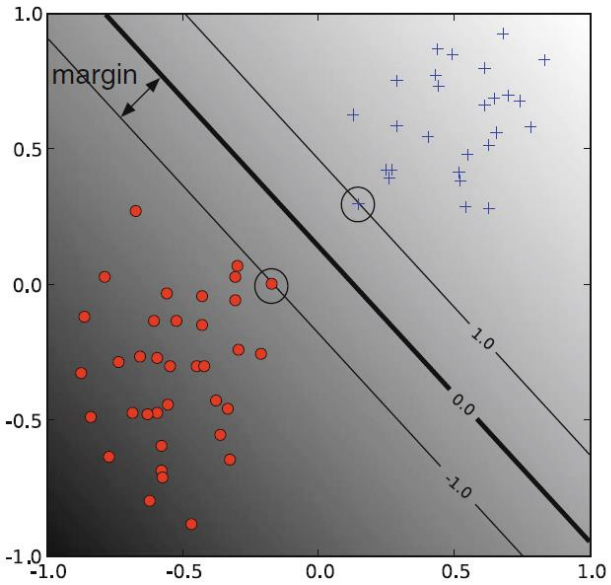

Figura 6.3 Margen en SVM. Encerrados en círculos aparecen los vectores de soporte [Ben-Hur & Weston, 2010].

En SVM, el concepto de margen es de gran importancia y se define como la menor distancia entre la frontera de decisión y cualquiera de los puntos de los datos de

entrenamiento. Luego, la frontera de decisión se escoge de manera tal que se maximice el margen (Figura 6.3).

6.1.3. Métodos de evaluación: validación cruzada de *k* hojas

En las tareas de clasificación, existen varias técnicas o métodos para obtener la efectividad del algoritmo de clasificación. Una de ellas es utilizar un conjunto de datos de entrenamiento para construir el modelo y luego probar la efectividad sobre el mismo conjunto de datos. Esta técnica puede arrojar resultados no confiables pues el modelo se ha construido a la medida de este conjunto de datos y puede no funcionar adecuadamente para otro conjunto de datos suministrado.

Una estrategia para evitar este problema es separar el conjunto de datos de modo que una parte permanece como de clasificación desconocida y es usada para probar la efectividad del clasificador. Una versión mejorada de este procedimiento es la validación cruzada.

La validación cruzada consiste en dividir el conjunto de datos y realizar entrenamientos sucesivos donde las partes son empleadas como datos de entrenamiento y de prueba alternativamente [Kohavi, 1995]. La Figura 6.4 muestra la idea general de este procedimiento.

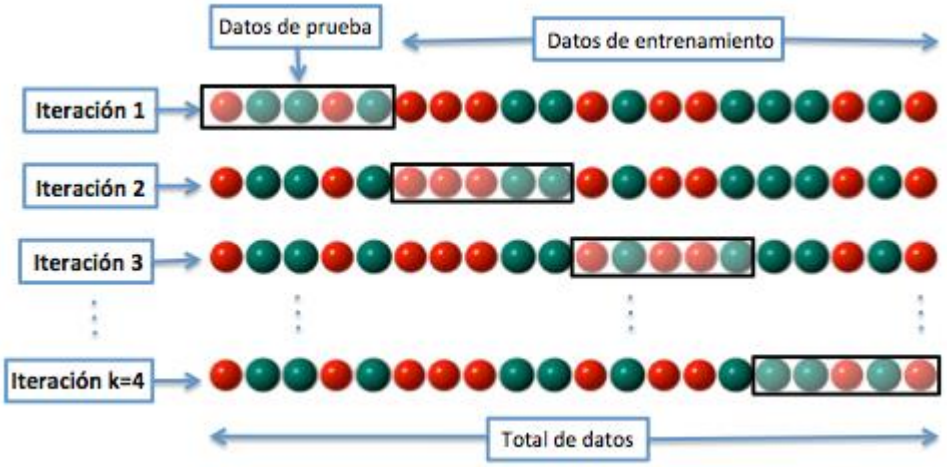

Figura 6.4 Validación cruzada de 4 hojas [Wikipedia, 2014].

En este tipo de validación el conjunto de datos D es dividido en k subconjuntos mutuamente exclusivos (las hojas) de aproximadamente el mismo tamaño. Este conjunto de datos es entrenado y probado k veces. En cada iteración $t \in \{1,2,\dots,k\}$, el entrenamiento se realiza sobre $D \backslash D_t$ y las pruebas se realizan sobre D_t. La estimación

de acierto está data entonces por el promedio de los aciertos en cada una de las iteraciones. Cuando $k = n$, donde n es la cantidad de instancias en el conjunto de datos, la validación se conoce como leave-one-out [Kohavi, 1995].

6.2. Calibración de parámetros para el algoritmo SVM

Para el uso del algoritmo SVM es necesario elegir el tipo de núcleo que se va a emplear y calibrar los parámetros en función de esta elección. A continuación se detalla el procedimiento aplicado en el presente trabajo de investigación.

6.2.1. Elección del núcleo

Las máquinas de soporte vectorial pertenecen a la categoría de métodos de núcleo, esto significa que el algoritmo manipula los datos solamente mediante productos punto [Ben-Hur & Weston, 2010]. Estos pueden ser reemplazados por funciones de núcleo, que computan estos productos punto en espacios de características de mayor dimensión [Ben-Hur & Weston, 2010].

La elección del núcleo tiene un gran impacto en el funcionamiento del algoritmo y va a depender de las características de los datos de cada problema: la cantidad de datos, su dimensionalidad, etc. Los núcleos más usados en la literatura son [Chapelle, 1998]:

- Lineal
- Polinomial
- Gaussiano o Función de Base Radial (RBF)
- Sigmoide

Los parámetros de cada núcleo también tienen efecto en los resultados que se obtengan. Así, por ejemplo, se tiene el grado del núcleo polinomial. El núcleo polinomial de menor grado es el núcleo lineal, pero este puede no ser suficiente si existe una correspondencia no lineal entre algunas de las características. Un ejemplo de esto puede observarse en la Figura 6.5, donde mientras mayor es el grado del núcleo polinomial, se obtiene una frontera de decisión más flexible [Ben-Hur & Weston, 2010].

En general, según se enuncia en [Chih-Wei, Chih-Chung, & Chih-Jen, 2010], el núcleo RBF es una buena elección. A diferencia del núcleo lineal, el RBF puede manejar relaciones no lineales entre los atributos y la clase de las entidades. Además, el núcleo lineal y sigmoide pueden comportarse de manera similar al núcleo RBF en cuanto a

porcentaje de acierto, bajo algunos valores de los parámetros [Chih-Wei, Chih-Chung, & Chih-Jen, 2010].

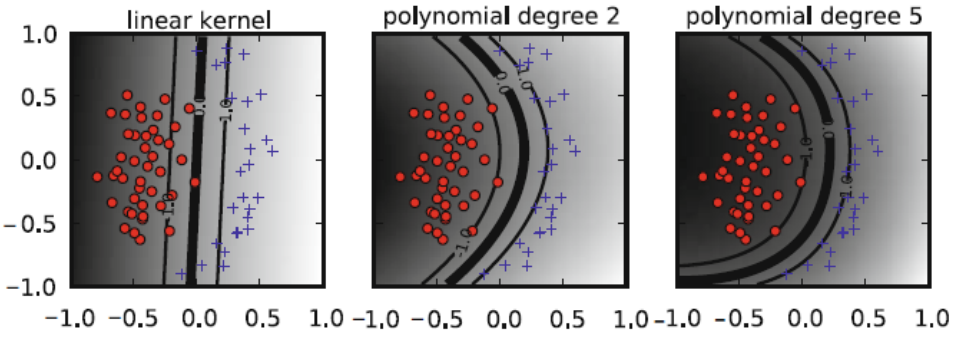

Figura 6.5 Efectos del grado en el núcleo polinomial [Ben-Hur & Weston, 2010].

Otro elemento que se consideró fue la dificultad de la elección de los parámetros en función de cada núcleo. En este sentido, el núcleo polinomial tiene más parámetros que el núcleo RBF y esto dificulta las tareas de calibración [Chih-Wei, Chih-Chung, & Chih-Jen, 2010].

De manera preliminar, se experimentó con la función Classify de Mathematica [Wolfram, 2014] utilizando como método SVM y como tipo de núcleo el Lineal, el Polinomial y el RBF. El conjunto de datos fue dividido de manera aleatoria en 300 entidades para entrenamiento y el resto para pruebas. Para obtener los porcentajes de acierto se utilizó la función ClassifierMeasurements de Mathematica [Wolfram, 2014]. Los porcentajes de acierto obtenidos se muestran en la Tabla 6.1 y refuerzan el criterio de seleccionar el núcleo RBF.

Tabla 6.1 Porcentaje de acierto con distintos núcleos.

Núcleo Lineal	77.05%
Núcleo Polinomial	63.11%
Núcleo RBF	81.97%

6.2.2. Calibración de parámetros

En el algoritmo SVM con núcleo RBF hay dos parámetros: el factor de penalidad C y el parámetro de núcleo γ [Chih-Wei, Chih-Chung, & Chih-Jen, 2010]. La estrategia utilizada para la calibración fue una búsqueda en grilla utilizando validación cruzada, como se recomienda en [Chih-Wei, Chih-Chung, & Chih-Jen, 2010]. De esta manera se

experimentan con diferentes pares (C, γ) y se selecciona la combinación que mejores resultados haya tenido. Además, se sugiere incrementar los valores de los parámetros de manera exponencial, para escapar a los óptimos locales [Chih-Wei, Chih-Chung, & Chih-Jen, 2010].

La herramienta Mathematica, permitió realizar esta exploración sucesiva de variantes y mostrar los resultados en gráficos bidimensionales, donde se observó los intervalos de parámetros con que mejores resultados se obtenían. Se hicieron dos iteraciones, una primera iteración, con un rango numérico más grande y espaciado para los parámetros (C, γ) y una segunda iteración para reducir los rangos de búsqueda y detallar más en los valores numéricos.

Para la primera iteración se utilizó la siguiente configuración de parámetros: $C = 2^{-3}, 2^{-1}, \dots, 2^{15}$ y $\gamma = 2^{-15}, 2^{-13}, \dots, 2^{3}$. Los resultados obtenidos se muestran en la Figura 6.6. A partir del análisis de los resultados de la primera iteración, se seleccionaron los parámetros de la segunda iteración de la siguiente forma: $C = 2^{-1}, 2^{-0.75}, \dots, 2^{1}$, $\gamma = 2^{-3}, 2^{-2.75}, \dots, 2^{-1}$. Los resultados se muestran en la Figura 6.7. De este análisis se concluyó utilizar los valores de $C = 2^{0.75}$ y $\gamma = 2^{-2.75}$.

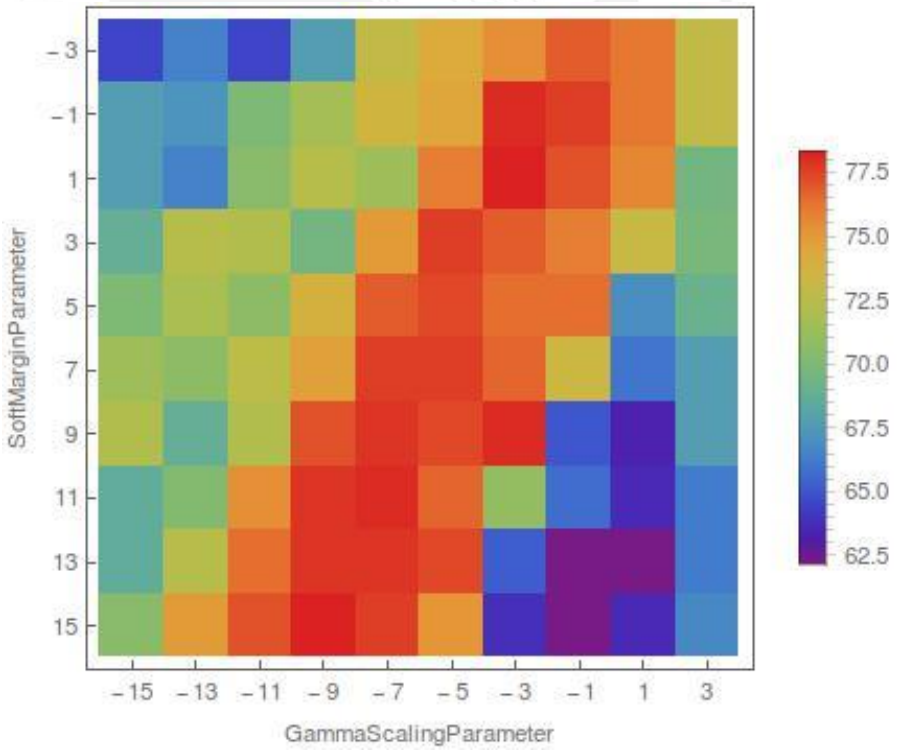

Figura 6.6 Calibración de parámetros, primera iteración.

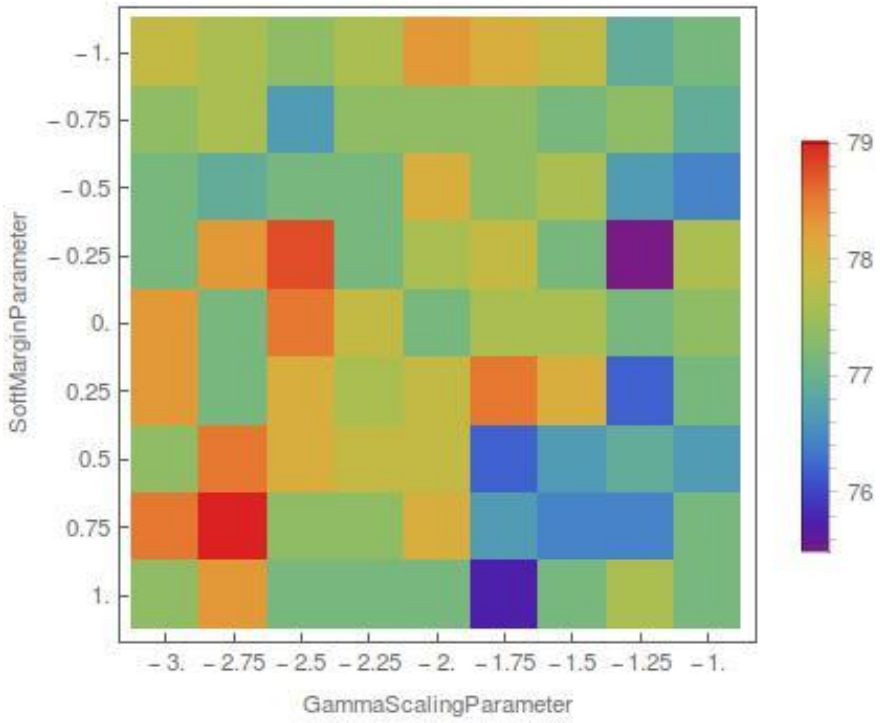

Figura 6.7 Calibración de parámetros, segunda iteración.

6.3. Aplicación del algoritmo SVM usando Mathematica

La función Classify [Wolfram, 2014] genera una función clasificadora, o modelo de clasificación a partir de un conjunto de datos de entrenamiento suministrado. Además, es posible especificar el método de clasificación de entre un grupo de 6 algoritmos que ya se encuentran implementados, así como otras opciones relacionadas tanto al proceso de aprendizaje como al algoritmo seleccionado propiamente.

En este sentido, existe la función ClassifierInformation [Wolfram, 2014], la cual genera un reporte sobre una función clasificadora dada. Este reporte incluye información general sobre el método utilizado para la clasificación, número de clases y atributos, número de instancias, entre otros.

La función que se muestra en la Figura 6.8 devuelve el porcentaje de acierto de la clasificación que se realiza dividiendo de manera aleatoria los datos en conjunto de entrenamiento y conjunto de pruebas: 300 entidades se utilizan durante el entrenamiento o construcción de la función clasificadora, el resto de las entidades forman parte del conjunto de pruebas. Esta función tiene en cuenta la selección de los parámetros realizada previamente. Luego de aplicar sucesivamente 100 veces esta función clasificadora al conjunto de datos, se obtuvo una tasa de acierto promedio de 77.6%. El

histograma de la Figura 6.9 muestra información más detallada sobre las tasas de acierto obtenidas.

```
ClassifierEvaluator (inputData)
  mean = 0
  FOR i = 1 to 100
    trainingSet = RandomSample(inputData, 300)
    testSet = Complement(inputData, trainingSet)

    svmClassifier = Classify(trainingSet, Method -> {SupportVectorMachine,
                                Kernel -> RBF, GammaScalingParameter -> 2^-2.75,
                                SoftMarginParameter -> 2^0.75})
    mean = mean + ClassifierMeasurements(svmClassifier, testSet).Accuracy

  return mean / 100;
END
```

Figura 6.8 Función que evalúa la tasa de acierto del clasificador.

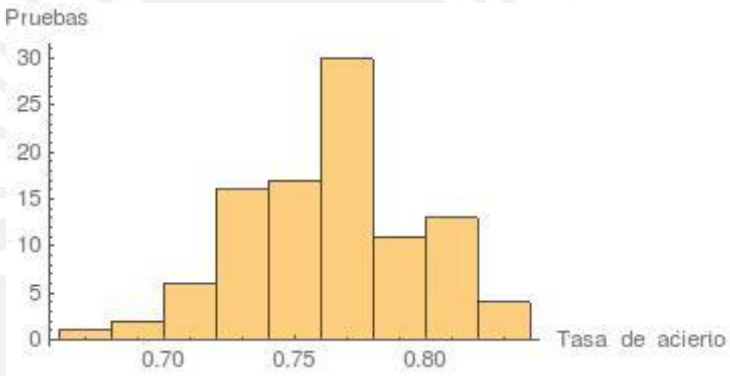

Figura 6.9 Histograma de las tasas de acierto obtenidas en las 100 iteraciones realizadas.

6.3.1. Matriz de confusión

La matriz de confusión es una tabla que muestra una relación de la cantidad de instancias clasificadas por cada clase. Si se considera un problema de clasificación con tres clases C_1, C_2 y C_3, la matriz de confusión sería la que se ilustra en la Figura 6.10. Los valores en la diagonal representan el número de instancias correctamente clasificadas por cada clase, mientras que el resto de los valores representan instancias clasificadas erróneamente [Roiger & Geatz, 2003]. Por ejemplo, el valor C_{23} indica la cantidad de instancias que eran de tipo C_2 pero que fueron clasificadas erróneamente como de tipo C_3.

	C_1	C_2	C_3
C_1	C_{11}	C_{12}	C_{13}
C_2	C_{21}	C_{22}	C_{23}
C_3	C_{31}	C_{32}	C_{33}

Figura 6.10 Ejemplo de matriz de confusión [Roiger & Geatz, 2003].

La Figura 6.11 muestra la matriz de confusión correspondiente a la clasificación de las células espermáticas que se realizó en la sección anterior. Del conjunto de pruebas de 122 células, 76 eran anormales y 46 normales. La matriz generada muestra que 31 células con cabezas morfológicamente normales fueron clasificadas correctamente y 15 de ellas fueron clasificadas erróneamente. Por otra parte, de las células con cabezas morfológicamente anormales, 68 fueron clasificadas correctamente y solo 8 fueron clasificadas erróneamente.

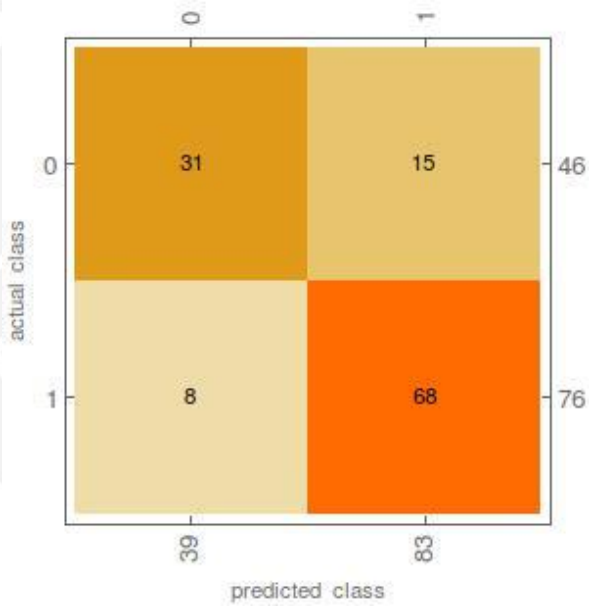

Figura 6.11 Matriz de confusión.

7. ANÁLISIS COMPARATIVO DE LA CARACTERIZACIÓN MORFOLÓGICA

En esta sección se muestra un análisis comparativo entre los valores de las características morfológicas obtenidas para las células normales según las mediciones realizadas y los valores recomendados por la OMS. Para ello, se tuvo en cuenta el largo y ancho de la cabeza de las células. Las células consideradas como normales, son aquellas que fueron clasificadas como tal por los especialistas.

La Figura 7.1 muestra un histograma de los valores del largo de la cabeza de las células normales, medidos en micrómetros. Con respecto a esta característica, la OMS establece que la media debe ser de 4.1 micrómetros, con un intervalo de confianza del 95% entre 3.7 y 4.7 micrómetros. En el caso de la muestra que se ha empleado en la presente investigación, la media del largo es de 4.58 micrómetros y el 59% de los valores medidos para las 155 células normales se encuentran en el intervalo de confianza antes mencionado.

Además, se tomaron 100 muestras aleatorias de 50 entidades de entre las 155 células normales, y se analizó el largo medio para estas muestras. El histograma de la Figura 7.2 muestra los resultados de este análisis, donde el 96% de los valores calculados se encontraron dentro del intervalo de confianza establecido por la OMS.

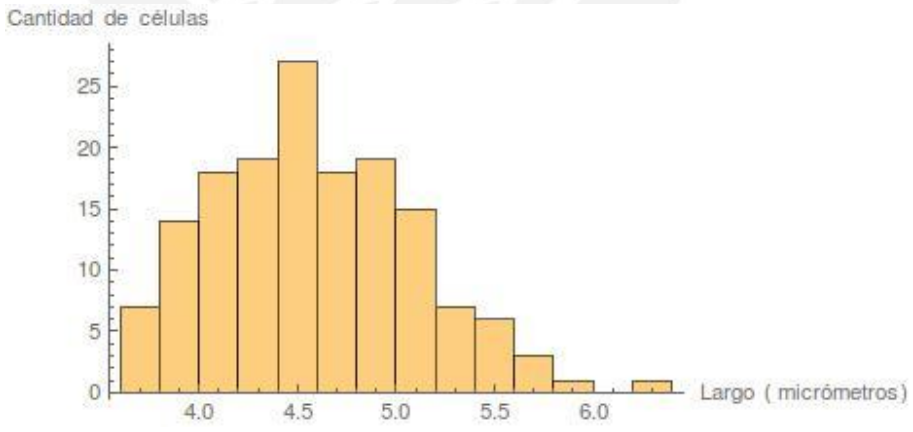

Figura 7.1 Histograma de los valores de largo medidos para las 155 células normales.

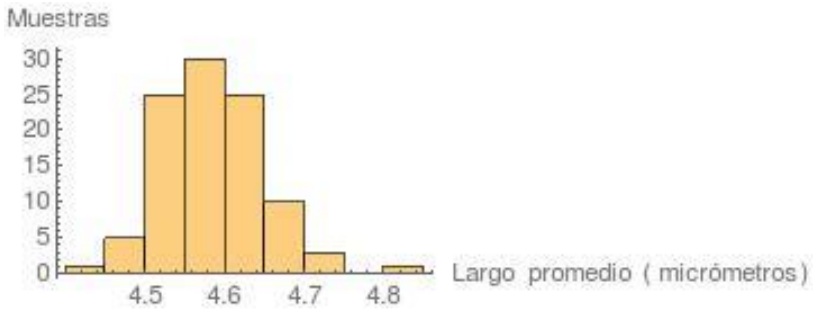

Figura 7.2 Histograma de las medias del largo, tomado de 100 muestras aleatorias de 50 células.

De manera similar se realizó el análisis del ancho. La Figura 7.3 muestra un histograma de los valores del ancho de la cabeza de las células normales, medidos en micrómetros. En este caso, la OMS establece que la media debe ser de 2.8 micrómetros, con un intervalo de confianza del 95% entre 2.5 y 3.2 micrómetros. En el caso de la muestra que se ha empleado en la presente investigación, la media del ancho es de 3.05 micrómetros y el 75% de los valores medidos para las 155 células normales se encuentran en el intervalo de confianza antes mencionado.

Análogamente a lo realizado con el análisis del largo, se tomaron 100 muestras aleatorias de 50 entidades de entre las 155 células normales, y se analizó el ancho medio. El histograma de la Figura 7.4 muestra los resultados de este análisis, donde el 100% de los valores calculados se encontraron dentro del intervalo de confianza establecido por la OMS.

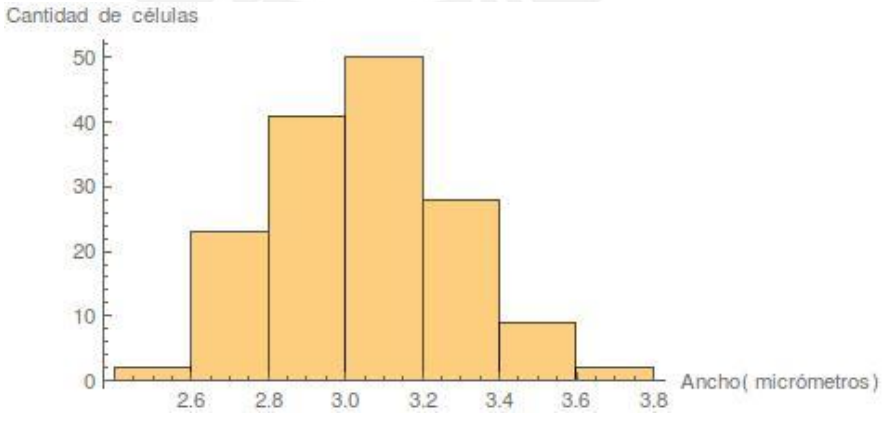

Figura 7.3 Histograma de los valores de largo medidos para las 155 células normales

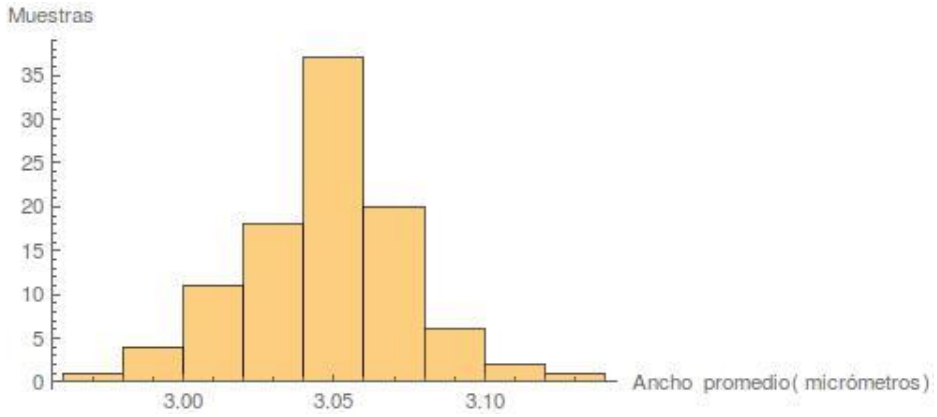

Figura 7.4 Histograma de las medias del ancho, tomado de 100 muestras aleatorias de 50 células.

8. DISCUSIÓN DE LOS RESULTADOS

El modelo algorítmico desarrollado provee un conjunto de pasos a seguir para lograr la caracterización y clasificación morfológica de cabezas de células espermáticas a partir del análisis de micrografías digitales. En esta sección se analizan los resultados obtenidos en cada etapa del modelo computacional propuesto.

En general, puede decirse que los resultados obtenidos son buenos y que el modelo desarrollado es extensible, considerando que se conoce el factor de conversión de pixeles a micrómetros de las imágenes que se quieren analizar.

8.1. Procesamiento de imágenes microscópicas de células espermáticas

En el caso de las imágenes extraídas del manual de la OMS, el proceso de binarización aplicado permitió identificar la cabeza de las células espermáticas en las imágenes microscópicas de las muestras. Del total de 226 células originales, se pudieron identificar las cabezas de 221 de ellas, o sea un 97.8%.

8.2. Extracción de características morfológicas

En cuanto a la extracción automática de características a partir de las imágenes binarizadas, se tiene que del total de 466 células iniciales (de las dos fuentes), se identificaron 422 células, o sea, un 91.5%. Si se compara este resultado con los porcentajes obtenidos en otros estudios similares, puede considerarse que el resultado obtenido es bueno. Por ejemplo, la herramienta computarizada desarrollada por [Carrillo, Villareal, Sotaquirá, Goelkel, & Gutiérrez, 2007] tiene un porcentaje de células identificadas del 89.5% como resultado de un proceso de segmentación.

Del total de 461 células iniciales, 160 eran normales y 301 anormales y 5 no estaban clasificadas. El 96.9% de las células normales fueron detectadas de manera automática, mientras que el porcentaje de células anormales detectadas fue de 88.7%. Esto se explica pues algunas de estas células clasificadas como anormales, tienen formas irregulares no ovaladas o una coloración no estándar que imposibilitaron la detección por parte de la herramienta automática.

8.3. Clasificación morfológica

En cuanto a la clasificación morfológica de la cabeza de las células espermáticas, la tasa de acierto obtenida es de aproximadamente 77.6%. Sin embargo, el 86% de

cabezas con una morfología anormal fueron clasificadas como tal, lo cual permite hacer un descarte de células anormales con una probabilidad de acierto alta. El hecho de que las cabezas normales hayan sido más difíciles de clasificar correctamente de manera automática es consecuente con el fenómeno de la heterogeneidad de las células normales y la dificultad de encontrar un patrón lo suficientemente estándar para identificar a las mismas [World Health Organization, 2010].

La comparación del resultado obtenido en la clasificación con los resultados obtenidos en las investigaciones consultadas en el estado del arte no puede realizarse de manera directa. La razón es que estas investigaciones contemplan la clasificación de la célula completa, teniendo en cuenta aspectos de la pieza media, la cola, entre otros. Solo como marco de referencia, se incluye la Figura 8.1 donde aparecen los porcentajes de acierto obtenidos usando diversos algoritmos de clasificación para el caso de una investigación donde se utilizaron los niveles de gris y curvas de contorno para caracterizar las células espermáticas [Tseng, et al., 2013]. Se puede observar que los mejores resultados se obtuvieron utilizando un algoritmo de clasificación basado en SVM (88.90% de acierto), pero en promedio el porcentaje de acierto fue de 76.88%, resultado similar al obtenido con el modelo propuesto en la presente investigación (77.6%).

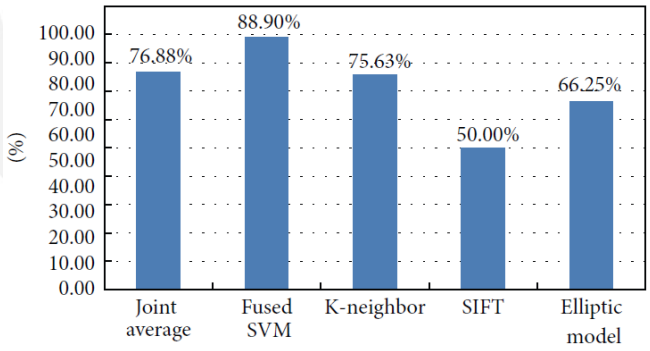

Figura 8.1 Comparación de métodos de clasificación [Tseng, et al., 2013].

8.4. Análisis comparativo de la caracterización morfológica

En el capítulo 7 se realizó un análisis de los valores de largo y ancho de la cabeza de las células espermáticas obtenidos en la caracterización morfológica realizada. Teniendo en cuenta las medias de las muestras aleatorias tomadas del conjunto de las 155 células normales analizadas, se observó que las mismas se encontraban en el intervalo de confianza propuesto por la OMS. El hecho de que algunas mediciones de manera

independiente se encontraran fuera de estos intervalos se explica por los siguientes factores que se mencionan en el manual de la OMS [World Health Organization, 2010]:

- Influencia del procedimiento realizado para la preparación de las muestras.
- Errores de aproximación en la selección del factor de conversión de pixeles a micrómetros.
- Los valores propuestos por la OMS han sido tomados de manera experimental, a partir de un conjunto finito de imágenes.
- La clasificación de las células contenidas en estas imágenes en normales o anormales no se realizó mediante métodos computarizados, sino de manera manual de acuerdo al criterio de un especialista.

9. OBSERVACIONES, CONCLUSIONES Y RECOMENDACIONES

En el presente capítulo se presentan las observaciones que se han podido levantar durante el proyecto de investigación. También se presentan las conclusiones del trabajo y finalmente las recomendaciones finales y trabajo futuro que queda propuesto.

9.1. Observaciones

Durante el desarrollo del presente proyecto de investigación se ha observado que para garantizar el éxito no solo de este proyecto en particular sino de cualquier proyecto similar, son necesarios un conjunto de recursos adicionales al propio esfuerzo del investigador. En este caso, la disponibilidad de una base de datos de microimágenes de células espermáticas, los patrones de anormalidad y normalidad definidos por los especialistas y la validación oportuna de los resultados obtenidos en cada etapa por parte de los especialistas en la materia fue de gran importancia.

Se hace necesario notar que las imágenes que se han empleado como base de la investigación se han obtenido en ambientes controlados, por lo que en algunos casos, los resultados de la investigación no se podrán extender de manera directa a otros conjuntos de datos. En este sentido, los patrones de anormalidad / normalidad han sido definidos por especialistas por lo que esto también aporta un factor subjetividad a la investigación y al modelo automático propuesto.

La herramienta Mathematica de Wolfram [Wolfram, 2014] aportó gran valor al trabajo desarrollado al jugar un papel fundamental en la investigación, desde el procesamiento de las imágenes hasta la clasificación. Todo esto gracias a la gran cantidad de funciones implementadas en esta herramienta, el fácil manejo de datos en distintos formatos, la posibilidad de automatizar operaciones y de visualizar los resultados en todo tipo de gráficos.

9.2. Conclusiones

Como resultado de la presente investigación se obtuvo un modelo algorítmico para la identificación de células espermáticas mediante el análisis automático de micrografías digitales. Dicho modelo permite la caracterización morfológica de la cabeza de las células así como su clasificación en normales o anormales, partiendo de un conjunto de datos de entrenamiento evaluado por especialistas.

El procesamiento de las imágenes permitió identificar las cabezas de las células presentes en ellas a través de una imagen binarizada. A partir de dicha imagen, fue posible extraer de manera automática las características morfológicas de la cabeza de las células, como por ejemplo el largo, ancho, área, perímetro, entre otras.

La aplicación de técnicas de estandarización y del algoritmo PCA, permitió expresar el conjunto de características extraídas de las microimágenes, de modo que se evidenciaran las correlaciones entre dichas características y pusieran seleccionarse aquellas que mejor identificaran al conjunto de datos. Dichos procedimientos llevaron a que se obtuvieran mejores resultados en el proceso de clasificación de las cabezas de las células espermáticas en normales o anormales, de acuerdo a su morfología.

El algoritmo SVM y los valores de los parámetros seleccionados en el proceso de calibración, como el tipo de núcleo y el parámetro de penalidad, hicieron posible que se obtuviera una tasa de acierto de aproximadamente 77.6%, lo cual de acuerdo a las características del problema y a los estudios citados en el estado del arte, es un valor aceptable.

La predicción de normalidad o anormalidad del modelo desarrollado es considerada un resultado parcial en el análisis de la infertilidad masculina a través de la morfología de las células espermáticas, ya que al mismo habría que sumar el análisis de la pieza media y de la cola, así como de otros elementos de la cabeza, como el porcentaje de acrosoma. Sin embargo, es una ventaja contar con este resultado, pues se puede realizar un descarte de aquellas células que no tienen una cabeza con forma normal, y continuar el análisis con aquellas que sí, lo cual reduce considerablemente el número de entidades a tratar, dado que el porcentaje de células normales en una muestra ya sea de un hombre fértil o infértil es generalmente inferior al 30% [World Health Organization, 2010].

9.3. Recomendaciones y trabajo futuro

La primera recomendación que se propone es ampliar y estandarizar la fuente de micrografías digitales de células espermáticas. Mientras más datos se tengan para entrenar al clasificador y mejor sea la calidad de los mismos, se obtendrán mejores resultados.

Como extensión al modelo propuesto se podría modificar el método de clasificación para que además de proveer información sobre la normalidad o no de la cabeza de las células espermáticas, pueda distinguir tipo de anomalía en el caso de las células anormales. Esta tarea se facilitaría si se amplía y estandariza la fuente de imágenes, como se indica en el punto anterior.

Además, se propone mejorar los métodos de identificación de la cabeza de las células espermáticas de modo que también se pueda obtener información sobre el acrosoma y el núcleo. Esto haría posible que se pudiera realizar una clasificación completa de la cabeza de la célula y que se puedan detectar otro tipo de anomalías, como por ejemplo la ausencia de acrosoma.

Otra recomendación sería emplear la simetría como una característica de la cabeza de la célula, expresada numéricamente a tener en cuenta en su caracterización y clasificación. Con el objetivo de completar la clasificación morfológica de la célula, se podría extender el modelo al análisis de la pieza media y la cola. De esta forma se podría dar el diagnóstico general del análisis morfológico.

En un futuro, sería interesante unir a este modelo el análisis de motilidad y conteo de células espermáticas, de modo que se pueda generar un espermiograma completo de manera automática.

BIBLIOGRAFÍA

Amann, R. P., & Waberski, D. (2014). Computer-assisted sperm analysis (CASA): Capabilities and potential developments. *Theriogenology*, 5-17.

Amann, R. P., & Waberski, D. (2014). Computer-assisted sperm analysis (CASA): Capabilities and potential developments. *Theriogenology*, 5-17.

American Society for Reproductive Medicine. (2006). Significance of sperm characteristics in the evaluation of male infertility. *Fertility and Sterility, 85*(3), 629-634.

Auger, J. (2010). Assesing human sperm morphology: top models, underdogs or biometrics? *Asian Journal of Andrology*, 36-46.

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. In O. Carugo, & F. Eisenhaber, *Data Mining Techniques for the Life Sciences, Methods in Molecular Biology.* Humana Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York, USA: Springer.

Butts, I., Ward, M., Litvak, M., Pitcher, T., Alavi, S., Trippel, E., & Rideout, R. (2011). Automated sperm head morphology analyzer for open-source software. *Theriogenology, 76*, 1756-1761.

Carrillo, H., Villareal, J., Sotaquirá, M., Goelkel, Á., & Gutiérrez, R. (2007). *A Computer Aided Tool for the Assessment of Human Sperm Morphology.* Barranquilla, Colombia: IEEE.

Chang, V., Saavedra, J. M., Castañeda, V., Sarabia, L., Hitschfeld, N., & Hartel, S. (2014). Gold-standard and improved framework for sperm head segmentation. *Computer Methods and Programs in Biomedicine*, 1-13.

Chapelle, O. (1998). *Support Vector Machines for Image Classification.* Redbank, NJ, USA: Image Processing Research Department, AT&T.

Chia, S., Tay, S., & Lim, S. (1998). What constitutes a normal seminal analysis? Semen parameters of 243 fertile man. *Human Reproduction, 13*(12), 3394-3398.

Chih-Wei, H., Chih-Chung, C., & Chih-Jen, L. (2010). *A Practical Guide to Support Vector Classification.* Taiwan: Department of Computer Science, National Taiwan Univeristy.

Cooper, T. G., & Noonan, E. (2010). World Health Organization reference values for human semen characteristics. *Human Reproduction Update, 16*(3), 231-245.

Esteves, S. C., Zini, A., Aziz, N., Alvarez, J. G., Sabanegh, E. S., & Agarwal, A. (2012). Critical Appraisal of World Health Organization's New Reference Values for Human Semen Characteristics and Effect on Diagnosis and Treatment of Subfertile Man. *Urology*, 16-22.

GIMP. (2014). *GNU Image Manipulation Program*. From GIMP: http://www.gimp.org/

Hamilton-Thorne. (2014, 04 02). *CEROS II Clinical*. From Innivations to Rely in Hamilton-Thorne: http://www.hamiltonthorne.com/index.php/products/clinical-casa-products/ceros-ii-clinical

Hatcher, L. (1994). Principal Component Analysis. In L. Hatcher, *A Step-by-step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling* (pp. 1-56). SAS Institute.

Kohavi, R. (1995). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* Stanford, CA: Computer Science Department, Stanford University.

Lu, J. C., Huang, Y. F., & Lu, N. Q. (2013). Computer-aided sperm analysis: paso, present and future. *Andrologia, xx*, 1-10.

Nikolettos, N., Kupker, W., Demirel, C., Schopper, B., Blasig, C., R., S., . . . Al-Hasani, S. (1999). Fertilization potential of spermatozoa with abnormal morphology. *Human Reproduction*, 47-70.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 62-66.

Pepper-Yowell, A. R. (2011). *Thesis: The use of Computer Assisted Semen Analysis to predict fertility in Holstein Bulls.* Fort Collins, Colorado: Department of Animal Sciences, Colorado State University.

Rijsselaere, T., Van Soom, A., Hoflack, G., Maes, D., & de Kruif, A. (2004). *Automated sperm morphometry and morphology analysis of canine semen by the Hamilton-Thorne analyser.* Merelbeke, Belgium: Elsevier.

Roiger, R., & Geatz, M. (2003). Data Mining: A Closer Look. In R. Roiger, & M. Geatz, *Data Mining, A Tutorial-based Primer* (pp. 33-65). Pearson Edducation.

Shlens, J. (2003, March 25). A Tutorial on Principal Component Analysis.

Smith, L. I. (2002). *A tutorial on Principal Components Analysis.* USA: Cornell University.

Tseng, K.-K., Li, Y., Hsu, C.-Y., Huang, H.-N., Zhao, M., & Ding, M. (2013). *Computer-Assisted System with Multiple Feature Fused Support Vector Machine for Sperm Morphology Diagnosis.* China: Hindawi Publishing Corporation.

van Leeuwen, J. (2004). Approaches in Machine Learning. In *Algorithms in Ambient Intelligence.* Philips Research Book Series.

Wikipedia. (2014). *Otsu´s Method*. From Wikipedia: http://en.wikipedia.org/wiki/Otsu's_method

Wikipedia. (2014). *Validación Cruzada*. From Wikipedia: http://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

Wolfram. (2014). *Binarize*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/Binarize.html

Wolfram. (2014). *ClassifierInformation*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/ClassifierInformation.html

Wolfram. (2014). *ClassifierMeasurements*. From Wolfram Language & Systems Documentation Center: http://reference.wolfram.com/language/ref/ClassifierMeasurements.html

Wolfram. (2014). *Classify*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/Classify.html

Wolfram. (2014). *ComponentMeasurements*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/ComponentMeasurements.html

Wolfram. (2014). *Mathematica de Wolfram*. From Wolfram: http://www.wolfram.com/mathematica/

Wolfram. (2014). *PrincipalComponents*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/PrincipalComponents.html

Wolfram. (2014). *Standarize*. From Wolfram Language & System Documentation Center: http://reference.wolfram.com/language/ref/Standardize.html

World Health Organization. (1999). *WHO laboratory manual for the Examination and processing of human semen* (4th ed.). Geneva, Switzerland: World Health Organization.

World Health Organization. (2009). International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary of ART terminology. *Fertility and Sterility, 92*(5), 1520-1524.

World Health Organization. (2010). *WHO laboratory manual for the Examination and processing of human semen* (5th ed.). Geneva, Switzerland: World Health Organization.