



INAOE

“Análisis y clasificación de firmas espectrales utilizando técnicas de aprendizaje automático”

Por

Ana Patricia Avila Guzmán

**Tesis sometida como requisito parcial para obtener el grado de:
Maestría en Ciencias en el área de Ciencia y
Tecnología del Espacio
en el
Instituto Nacional de Astrofísica, Óptica y Electrónica.**

**Supervisada por:
Dra. Raquel Díaz Hernández**

**Sta. Ma. Tonantzintla, Cholula, Puebla
Febrero, 2019**

**©INAOE 2019
Derechos Reservados
El autor otorga al INAOE el permiso de reproducir y
distribuir copias de esta tesis en su totalidad o en partes.**



**Análisis y clasificación de firmas
espectrales utilizando técnicas de
aprendizaje automático**

Resumen

El estudio de firmas espectrales, hace posible la identificación de distintos objetos de la tierra y del cielo, presentes en imágenes digitales. Los elementos que en ella se encuentran la hacen poseer características particulares que contiene información sobre la materia con la que interaccionó; es análoga a una huella digital. Los investigadores estudian firmas espectrales, que representan propiedades físicas, químicas y biológicas, a través de su interacción con la radiación emite determinadas longitudes de onda del espectro electromagnético. Tiene múltiples aplicaciones en diferentes áreas, tales como la geociencia y la astronomía. En geociencia, los espectros son captados por satélites. Una vez que la radiación solar ha traspasado la atmósfera, cada tipo de superficie interactúa con la radiación de manera que absorbe longitudes de onda y refleja otras diferentes. En astronomía, los espectros estelares son captados por sensores. La radiación electromagnética que proviene de las estrellas en las que llega poca o ninguna radiación, es emitida en determinadas longitudes de onda del espectro y tiene líneas de absorción. En relación con el estudio de espectros estelares, el Instituto Nacional de Astrofísica, Óptica y Electrónica tiene a su resguardo el acervo de imágenes digitalizadas de las placas astronómicas que fueron tomadas con la Cámara Schmidt de Tonantzintla, desde 1944 hasta 1994, durante este período se realizaron observaciones del cielo, donde se muestreo todo el centro de la galaxia y uno de sus polos. El acervo de imágenes digitalizadas, se ha utilizado en otros trabajos; investigadores se han dedicado al estudio de espectros estelares, tanto de forma visual como automática. Con respecto a métodos automáticos, en el presente trabajo de tesis se propone un conjunto de datos, obtenido de algoritmos de extracción y selección de características que da como resultado la firma espectral de cada objeto estelar. De este modo, se realizó una clasificación de espectros estelares del conjunto de datos propuesto, utilizando técnicas de aprendizaje automático. El objetivo es clasificar el mayor número de espectros estelares e incrementar las clases y subclases reportadas en trabajos previos. Para finalizar con lo propuesto, se reportan los

resultados hasta un 90.32% de exactitud, para las clases y subclases principales de tipo espectral.

Abstract

The study of spectral signatures makes it possible to identify different objects of earth and sky, present in digital images. The elements that are in it make it have a particular feature, it is analogous to a fingerprint. Researcher's study its spectral signature, which is made up of the physical, chemical, biological and wavelength properties of electromagnetic energy. It has multiple applications in different areas, such as geoscience and astronomy. In geoscience, the spectra are captured by satellites, once the solar radiation has penetrated the atmosphere, each type of surface interacts with the radiation in a way that absorbs wavelengths and reflects different ones. In astronomy, the spectra of the stars are captured by sensors, the electromagnetic radiation that comes from the stars emits wavelengths of the spectrum and several absorption lines. In relation to the study of stellar spectra, the National Institute of Astrophysics, Optics, and Electronics has at its disposal the set of digitized images of the astronomical plates that were taken with Schmidt Camera of Tonantzintla, from 1944 to 1994, during this period observations, it sampled the entire center of the galaxy and one of its poles. The collection of digitized images has been used in other works; researchers have dedicated to the study of stellar spectra, visually and automatically. With respect to automatic methods, in the present thesis work, a set of data is proposed, obtained from algorithms of extraction and selection of feature which results in the spectral signature of each stellar object. In this way, classification of stellar spectra of the proposed data set was made, using machine learning. The objective is to classify the largest number of stellar spectra and increase the classes and subclasses reported in previous works. To finish with the proposed, the results are reported up to 90.32% accuracy, for the main classes and subclasses of spectral type.

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología CONACYT por la beca otorgada bajo el registro 656645, la cual sustentó mis estudios durante este período.

Al Instituto Nacional de Astrofísica, Óptica y Electrónica donde me he formado durante mis estudios de Maestría y donde he recibido apoyo de todo tipo. Doy gracias por tener la oportunidad de estar en esta Institución, donde tuve la dicha de conocer a excelentes personas, que me motivaron y alentaron para seguir en este camino de la Ciencia.

Gracias a la Dra. Raquel Díaz Hernández, quien ha supervisado los objetivos planteados, durante el desarrollo de este trabajo de tesis.

Gracias al Dr. Leopoldo Altamirano Robles por la atención brindada, aportando sus conocimientos, ideas y las herramientas necesarias para esta tesis.

Al equipo de trabajo del Laboratorio de Percepción Autónoma. M.C. Ariel, le agradezco por compartir sus conocimientos hacia mi persona, a la Ingeniera Noemí por la ayuda otorgada durante este trabajo de tesis, gracias.

Dedicatorias

Dedicado a mi familia, por ser mi mayor inspiración de lucha día con día, sin ellos nada de esto sería posible. Su apoyo incondicional y su amor han hecho de mí la persona que soy. La vida no viene con un manual de instrucciones, por suerte la mía vino con una excelente madre, exigente, seria, enojona, regañona, pero una persona honorable, invaluable, honesta, una guerrera que lucha día a día por el bienestar de su familia, la mejor consejera, la mejor amiga, la mejor de todas. El amor que me das, tú confianza, tú manera de apoyarme es incomparable, me has hecho fuerte, y me has enseñado tanto, así que dedicarte esta tesis es muy poco en comparación a todo lo que me has dado y has hecho por mí, pero con este trabajo de tesis, quiero retribuir a todo tu esfuerzo; gracias infinitas por ser mi madre, te amo.

A mi padre, si me lo preguntarás no cambiaría nada en ti, te quiero por ser tal y como eres, eres todo lo contrario a mamá y eso lo agradezco mucho, porque tú alegría, tranquilidad, amabilidad, espontaneidad, me motiva, me da fuerzas para lograr todo lo que me propongo, gracias por esta bonita familia.

Dedicado a mis hermanas, el amor que les tengo es infinito son mis mejores amigas estoy muy orgullosa de ustedes, están realizando sus sueños y luchan por conseguir todo lo que se proponen, su esfuerzo será recompensado. A Maythe por ser la hermana mayor que siempre quise, te amo mucho y agradezco todo el apoyo y el amor que me das.

A mi mejor amigo, a lo largo del tiempo me has enseñado tanto y admiro tu esfuerzo, entrega y dedicación, gracias por ser como eres y por apoyarme en todo momento, con mucho cariño para ti.

Índice general

Resumen	5
Abstract	7
Agradecimientos.....	9
Dedicatorias.....	11
Índice de figuras	16
Índice de tablas	17
Introducción.....	18
1.1 Planteamiento del Problema.....	20
1.2 Justificación.....	20
1.3 Objetivos.....	21
1.3.1 Objetivo general	21
1.3.2 Objetivos específicos	21
1.4 Alcances y Limitaciones.....	22
1.4.1 Alcances.....	22
1.4.2 Limitaciones	22
1.5 Estructura del documento	23
Marco Teórico.....	24
2.1 Firma espectral.....	24
2.2 Fundamentos teóricos	25
2.2.1 Tipos de espectros	26
2.2.2 Espectros estelares	28
2.2.3 Clasificación estelar automática	28
2.2.4 Clasificación espectral de Harvard	29
2.3 Procesamiento de imágenes digitales.....	31
2.3.1 Operaciones orientadas al punto	32
2.3.2 Operaciones espaciales	33
2.3.3 Operaciones basadas en regiones.....	34
2.3.4 Operaciones geométricas.....	35
2.4 Extractor de características WND-CHARM	35

2.4.1 Características de la imagen basadas en los coeficientes de Chebyshev	37
2.4.2 Características de la imagen basadas en la transformada de Chebyshev-Fourier	39
2.4.3 Características de la imagen basadas en momentos de Comb.	40
2.5 Selección de características algoritmo Relief-F	41
2.6 Técnicas de aprendizaje automático	42
2.6.1 Máquina de soporte vectorial	44
2.6.2 Árboles de decisión	44
2.6.3 Vecino más cercano	45
2.6.4 Análisis discriminante lineal	46
2.6.5 Redes Neuronales	46
2.7 Técnicas de evaluación	47
2.7.1 Matriz de confusión	48
2.7.2 Validación cruzada de K iteraciones	48
Trabajos Relacionados	49
3.1 Introducción del capítulo	49
3.2 Firmas espectrales	50
3.3 Espectros estelares	51
3.4 Extracción de características	53
3.5 Conclusión del capítulo	54
Trabajo desarrollado	55
4.1 Introducción del capítulo	55
4.2 Conjunto de datos propuesto	56
4.3 Acervo de imágenes astronómicas	59
4.4 Formato utilizado en el almacenamiento de las imágenes	59
4.5 Representación de las imágenes utilizadas en escala de grises	60
4.6 Procesamiento de imágenes	61
4.6.1 Mejoramiento de contraste	61
4.6.2 Realce de bordes	62
4.7 Segmentación de espectros	63
4.7.1 Rotación de la imagen	63
4.8 Extracción de características	64
4.8.1 Descriptores obtenidos	65

4.9 Selección de características.....	66
4.10 Técnicas de aprendizaje automático utilizadas.....	67
4.11 Clasificación de espectros estelares.....	68
4.12 Conclusión del capítulo	70
Experimentos y Resultados	71
5.1 Clasificación de espectros estelares.....	71
5.1.1 Fase I. Clasificación de cinco grupos de clases de tipos espectrales ..	72
5.1.2 Fase II. Clasificación con los coeficientes de Chebyshev y los momentos de Comb	75
5.1.3 Fase III. Clasificación de siete grupos de clases de tipos espectrales	78
5.2 Clasificación de subclases de espectros estelares.....	81
5.2.1 Fase IV. Clasificación de subclases de espectros estelares	81
5.3 Discusión de los resultados obtenidos	83
5.3.1 Fase I	83
5.3.2 Fase II.....	83
5.3.3 Fase III	84
5.3.4 Fase IV	84
5.3.5 Resultados de las fases propuestas	85
5.4 Comparación del método propuesto.....	86
Conclusiones	87
6.1 Conclusiones	87
6.2 Contribuciones.....	87
6.3 Publicaciones.....	88
6.4 Trabajo futuro	88
Referencias	89

Índice de figuras

Figura 1. Prisma de Newton, obtenida de [9]	25
Figura 2. Espectro electromagnético, obtenida de [10]	26
Figura 3. Espectro continuo, obtenida de [11].....	26
Figura 4. Espectro de líneas de emisión del sodio, obtenida de [11]	27
Figura 5. Espectro de líneas de absorción del sodio, obtenida de [11]	27
Figura 6. Espectro estelar.....	28
Figura 7. Clasificación espectral de Harvard, obtenida de [13].....	30
Figura 8. Etapas en el procesamiento de imágenes, obtenida de [15]	31
Figura 9. Características calculadas por el proceso WND-CHARM, obtenida de [25].....	37
Figura 10. Técnicas de aprendizaje automático: se incluyen el aprendizaje supervisado y el aprendizaje no supervisado.....	43
Figura 11. Red Neuronal Back Propagation, obtenida de [8]	47
Figura 12. Matriz de Confusión, obtenida de [37].....	48
Figura 13. Obtención del conjunto de datos propuesto.....	58
Figura 14. Placa tipo espectral.	59
Figura 15. Recorte de espectros estelares.....	63
Figura 16. Imagen rotada de espectros estelares	64
Figura 17. Imagen de un espectro estelar.....	64
Figura 18. Proceso de extracción de características.	65
Figura 19. Descriptores obtenidos de WND-CHARM	65
Figura 20. Proceso de selección de características	66
Figura 21. Método de clasificación de espectros estelares	69

Índice de tablas

Tabla 1. Técnicas de aprendizaje automático, obtenido de [10]	43
Tabla 2. Tabla comparativa.....	49
Tabla 3. Colección de imágenes de espectros estelares por tipo espectral.....	56
Tabla 4. Colección de imágenes de espectros estelares por subtipo espectral.....	57
Tabla 5. Conjunto de imágenes astronómicas.....	57
Tabla 6. Placas digitalizadas.....	61
Tabla 7. Mejoramiento de contraste	62
Tabla 8. Operador de Sobel.....	62
Tabla 9. Técnicas de aprendizaje automático.....	67
Tabla 10. Grupo con cinco clases de espectros estelares.....	72
Tabla 11. Matriz de Confusión Árbol de Decisión (5 clases)	73
Tabla 12. Matriz de Confusión KNN (5 clases).....	73
Tabla 13. Matriz de Confusión SVM (5 clases)	73
Tabla 14. Matriz de Confusión de Análisis Discriminante (5 clases)	73
Tabla 15. Matriz de Confusión de RN (5 clases).....	74
Tabla 16. Las 25 características más importantes obtenidas de WND-CHARM.....	75
Tabla 17. Matriz de Confusión Árbol de Decisión (5 clases).....	76
Tabla 18. Matriz de Confusión KNN (5 clases).....	76
Tabla 19. Matriz de Confusión SVM (5 clases)	76
Tabla 20. Matriz de Confusión Análisis Discriminante (5 clases).....	76
Tabla 21. Matriz de Confusión Redes Neuronales (5 clases).....	77
Tabla 22. Colección de imágenes de espectros estelares por tipo espectral.....	78
Tabla 23. Matriz de Confusión Árbol de Decisión (7 clases)	78
Tabla 24. Matriz de Confusión KNN (7 clases)	79
Tabla 25. Matriz de Confusión SVM (7 clases).....	79
Tabla 26. Matriz de Confusión Análisis Discriminante (7 clases)	79
Tabla 27. Matriz de Confusión Redes Neuronales (7 clases).....	80
Tabla 28. Colección de imágenes de espectros estelares por subtipo espectral	81
Tabla 29. Subclasificación de espectros estelares	82
Tabla 30. Resultados Primera Fase	83
Tabla 31. Resultados Segunda Fase	84
Tabla 32. Resultados Tercera Fase.....	84
Tabla 33. Resultados cuarta fase	85
Tabla 34. Resultados de las cuatro fases propuestas.....	85

Capítulo 1

Introducción

Las firmas espectrales representan propiedades físicas, químicas y biológicas, a través de su interacción con la radiación emite determinadas longitudes de onda del espectro electromagnético, los elementos que en ella se encuentran presentan un comportamiento diferente, el cual es captado por sensores. A esta característica particular de cada elemento, se le conoce como firma espectral, y es la base para procesos de clasificación, según Castilla Carrillo en [1]. En particular, si se observa el espectro electromagnético de una estrella, se notará que no es continuo, sino que le faltan tramos que muestran líneas oscuras, cada una de esas líneas indica que tipo de elemento químico hay en esa estrella. Así las líneas de absorción (las líneas oscuras) serán las firmas espectrales de los elementos, por ejemplo: el hidrógeno se encuentra en 7 líneas diferentes, y cada línea guarda su posición siempre en este arcoíris de la estrella que se conoce como espectro.

Actualmente, existe una clasificación espectral de las estrellas llamado “Sistema de Clasificación de Harvard”, el cual considera un conjunto de 7 clases principales de tipos espectrales estelares [O, B, A, F, G, K, M] siendo la de tipo O la estrella más caliente y la de tipo M la más fría debido a su temperatura y composición química, según Meadows en [2]. Frente a este tema de estudio, el Instituto Nacional de Astrofísica, Óptica y Electrónica tiene a su resguardo un gran número de imágenes digitalizadas de placas astronómicas. Los espectros estelares han sido capturados en estas placas que fueron tomadas con la Cámara Schmidt de Tonantzintla. La importancia histórica de esta colección radica en poseer las primeras capturas de múltiples regiones del cielo de esa época, donde se tomó todo el centro de la galaxia y uno de sus polos, mencionado en [3].

Dentro del análisis, un método que implementan los investigadores para realizar la clasificación de espectros estelares es utilizando el sentido de la vista, mediante su

visión y experiencia pueden determinar los tipos de espectros que aparecen en una imagen astronómica. La ventaja de este método es la exactitud con la que se realiza la clasificación y la desventaja es que toma bastante tiempo realizarlo.

De igual forma, otro método es a través del análisis de su firma espectral, el cual se realiza de forma automática con técnicas de *machine learning*; utilizando características propias de la imagen. La ventaja de este método es que se realiza de manera automática y el tiempo invertido es mínimo. Una desventaja, es que depende del aprendizaje adquirido para realizar una clasificación exacta. Esto quiere decir que al ser un método automático, se requieren de suficientes datos para realizar un buen entrenamiento. No obstante, las técnicas de aprendizaje automático reportan resultados con altos índices de clasificación.

Cabe mencionar, que en trabajos previos de la clasificación de espectros estelares, los autores Cruz Vega en [4] y Díaz Hernández en [5], utilizan el conjunto de datos del Catálogo de Pulkovo y sus enfoques de Espectrofotometría. De igual forma, clasificaron cinco de las clases [B, A, F, G, K] que presenta el Sistema de Clasificación de Harvard, según Meadows en [2], con índices de clasificación de alta precisión. Además implementaron métodos del Perceptrón Multicapa (MPL), Vecino más cercano (K-NN), Máquina de Soporte Vectorial (SVL) y Redes Neuronales (RN).

En el presente trabajo de tesis, se propone un conjunto de datos de firmas espectrales extraído del acervo de imágenes digitales, mediante algoritmos de extracción y selección de características. Con base en este conjunto, se logró una clasificación exacta y precisa de espectros estelares, utilizando cinco técnicas de aprendizaje automático. El objetivo fue clasificar la mayor cantidad de espectros estelares e incrementar el número de clases y subclases que se reportaron en trabajos previos. Para finalizar con lo propuesto, se muestran las tasas de exactitud para cinco y siete de las clases principales, así como también para cuatro sub clases de cinco clases principales, obteniendo hasta un 90.32% de exactitud en los resultados, lo que indica altos índices de clasificación.

1.1 Planteamiento del Problema

Actualmente, el Instituto Nacional de Astrofísica, Óptica y Electrónica tiene a su resguardo el acervo histórico de imágenes astronómicas digitalizadas para la clasificación de espectros estelares. La clasificación se realiza de forma visual por expertos en el área quienes mediante su visión y experiencia determinan la clase a la que pertenece el espectro que observan en la imagen. Este método requiere de dedicación exclusiva y de un arduo trabajo que se lleva a cabo por tiempo indeterminado. Esta situación motivó el desarrollo de métodos automáticos que realicen una clasificación de forma rápida y precisa.

1.2 Justificación

La clasificación estelar de una región del cielo es un proceso que requiere mucho tiempo debido a la gran cantidad de objetos estelares que se concentran en una sola imagen. Por estas razones, se requiere de un estudio exhaustivo para desarrollar este trabajo de tesis. A modo de conclusión, se necesita un método automático donde no se requiera de conocimientos previos por parte de un experto. De esta manera, surge un método eficaz para la clasificación estelar, de campo amplio para múltiples objetos y con posible aplicación a otra área como es Ciencias de la Tierra. En este sentido, el uso de técnicas automáticas acelerará todo el proceso sin necesidad de intervención humana, esto indicará la confiable extracción y selección de los datos. De acuerdo a las características análogas que posee un espectro, es más sencillo categorizarlo en clases y subclases, obteniendo una clasificación exacta y precisa.

1.3 Objetivos

1.3.1 Objetivo general

Identificar y clasificar espectros estelares extraídos del acervo digital de imágenes astronómicas tomadas con la Cámara Schmidt de Tonantzintla. Utilizando algoritmos de extracción y selección de características para crear el conjunto de datos de firmas espectrales. Mediante técnicas de aprendizaje automático se clasificará el conjunto de datos propuesto, se incrementará el número de clases y subclases con el fin de obtener una completa, exacta y precisa clasificación.

1.3.2 Objetivos específicos

- Identificar y extraer espectros estelares de imágenes astronómicas, mediante operaciones de procesamiento de imágenes para la confiable extracción de los datos.
- Utilizar algoritmos de extracción de características para obtener los valores más importantes de cada imagen.
- Crear un conjunto de datos de firmas espectrales mediante algoritmos de selección de características, este conjunto servirá de entrenamiento.
- Utilizar técnicas de aprendizaje automático para la clasificación de firmas espectrales de las imágenes de los espectros de las estrellas.
- Incrementar el número de clases y subclases permitidas de acuerdo a la Clasificación Espectral del Sistema de Harvard, con base en los trabajos previos reportados en la literatura.

1.4 Alcances y Limitaciones

1.4.1 Alcances

- La clasificación de espectros estelares contará con un conjunto de 3563 imágenes, obtenidas del acervo digital.
- Los algoritmos utilizados, ayudarán a obtener las características principales (firma espectral) que provienen directamente de la imagen de un espectro estelar.
- Los sistemas de aprendizaje automático, clasificarán el conjunto de datos propuesto con altos índices de exactitud y precisión, con base en los trabajos previos.
- La clasificación estelar detectará el mayor número posible de espectros de las clases y subclases con referencia al Sistema de Harvard.

1.4.2 Limitaciones

- Debido a la naturaleza de los espectros estelares, las clases de tipo O y M frecuentemente tienen entornos complicados por la falta de visibilidad en sus líneas de absorción que dificultan la medición de sus espectros; esto indica que se tendrán pocas muestras en comparación con las otras clases.

1.5 Estructura del documento

Para describir de manera detallada los conceptos necesarios para la realización de este trabajo de tesis, el documento presentado tiene la siguiente estructura:

Capítulo 1: Introducción, en este capítulo se explica el planteamiento del problema, la justificación; así mismo se puntualizan los objetivos, los alcances y limitaciones de este trabajo de tesis.

Capítulo 2: Marco teórico, se presentan los planteamientos teóricos que sustentan el desarrollo de esta tesis, se ofrecen conceptos generales de las técnicas utilizadas en el desarrollo del proyecto.

Capítulo 3: Trabajo relacionado, en este capítulo se presenta una descripción detallada de los trabajos relacionados a esta tesis con respecto a firmas espectrales, clasificación de espectros estelares y la extracción de características en imágenes digitales, finalmente se muestra la conclusión de este capítulo.

Capítulo 4: Trabajo desarrollado, en este capítulo se detalla el proceso que se llevó a cabo para la clasificación de espectros estelares, en primera instancia se muestra el conjunto de datos propuesto, el procesamiento de las imágenes utilizadas, el proceso de extracción y selección de características, posteriormente las técnicas de aprendizaje automático utilizadas y finalmente se exponen las conclusiones de este capítulo.

Capítulo 5: Experimentos y Resultados, en este capítulo se detallan los resultados obtenidos de las técnicas descritas en el Marco Teórico, se realizaron diferentes pruebas, logrando cumplir con los objetivos planteados en este trabajo de tesis.

Capítulo 6: Conclusiones, se muestran las conclusiones con base en los experimentos y resultados, se expone el trabajo que se pretende realizar a futuro.

En la sección final del documento se reportan las referencias consultadas.

Capítulo 2

Marco Teórico

Este capítulo, se encuentra organizado de la siguiente manera: en el primer subtema se hace referencia a la firma espectral. Posteriormente, se describen los fundamentos teóricos que enmarcan el desarrollo de este trabajo de tesis. En el siguiente subtema se explica el procesamiento de imágenes digitales, seguido de las operaciones utilizadas en el mismo contexto. Es importante destacar, el subtema de clasificación estelar automática, siendo el área en la que se orienta la aplicación del trabajo realizado. Para terminar, se describen los algoritmos de extracción y selección de características, y se detallan las técnicas de aprendizaje automático.

2.1 Firma espectral

El emisor de radiación más usual para imágenes de teledetección es el Sol. El Sol emite la radiación que incide, en primer lugar, en la atmósfera. Los gases presentes en la atmósfera, como el resto de la materia presente en el Universo, interactúa con la radiación, absorbiéndola, reflejándola o transmitiéndola [7]. Una vez que la radiación solar ha traspasado la atmósfera interactúa con la superficie terrestre, encontrándose con todo tipo de materiales diferentes, aguas dulces y saladas, tierras desnudas, nieve, zonas de vegetación densa, zonas de vegetación arbustiva, ciudades, etcétera. Cada tipo de superficie interactúa con la radiación de manera diferente, absorbiendo unas longitudes de onda muy concretas y reflejando otras diferentes en unas proporciones determinadas. Esta característica hace posible que se puedan identificar los distintos objetos tanto de la tierra como del cielo. Mediante experimentos de laboratorio se han podido caracterizar comportamientos de distintos objetos al recibir radiación.

A este comportamiento concreto de cada tipo de objeto se le llama firma o signatura espectral del mismo, según Hernández en [8].

2.2 Fundamentos teóricos

El presente estudio, es una revisión crítica de los elementos teóricos que sirven de referencia en esta investigación. Se establecen los conceptos relevantes, que se muestran a continuación.

Radiación electromagnética: La radiación electromagnética es un fenómeno de transporte de energía que puede ser únicamente observada por su interacción con la materia. La radiación electromagnética tiene componentes eléctricos y magnéticos, es afectada por estas propiedades con las cuales entran en contacto.

Espectro electromagnético: Proviene del latín *espectrum* y significa forma o aspecto [9]. Cuando se hace pasar la luz a través de un prisma óptico se produce el efecto llamado dispersión que consiste en la separación de las distintas longitudes de onda que forman el rayo incidente como se muestra en la figura 1.



Figura 1. Prisma de Newton, obtenida de [9]

Definición: En astronomía se le denomina espectro a la radiación electromagnética que emite o absorbe un objeto. Referido a un objeto se denomina espectro electromagnético o simplemente espectro a la radiación electromagnética que emite (espectro de emisión) o absorbe (espectro de absorción) una sustancia [9]. Dicho de otra forma, al conjunto de longitudes de onda que componen la radiación, se conoce como espectro electromagnético como se muestra en la figura 2. Aunque se trata de una sucesión continua de valores de longitud de onda, se suelen agrupar en bandas donde la radiación electromagnética manifiesta comportamientos similares, según Castilla en [6].

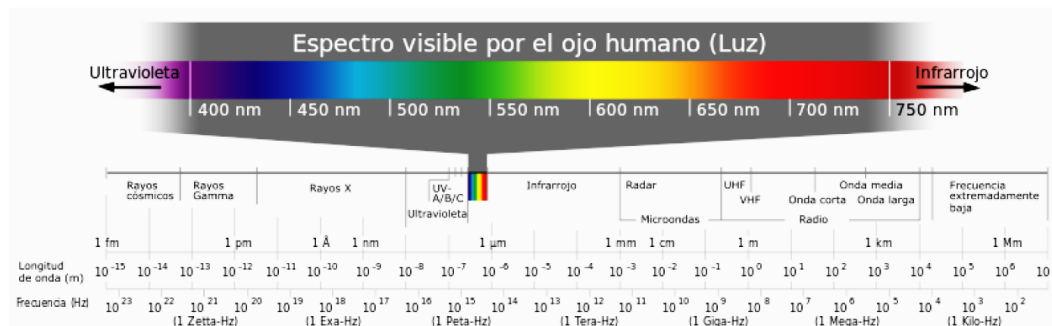


Figura 2. Espectro electromagnético, obtenida de [10]

2.2.1 Tipos de espectros

Dentro del análisis, los tipos de espectros se describen a continuación:

Espectro continuo: La luz blanca produce al descomponerla lo que se llama un espectro continuo [11], contiene el conjunto de colores que corresponde a la gama de longitudes de onda que la integran, como se muestra en la figura 3.



Figura 3. Espectro continuo, obtenida de [11]

Espectro con líneas de emisión: Sin embargo, los elementos químicos en estados gaseosos y sometidos a temperaturas elevadas producen espectros discontinuos en los que se aprecia un conjunto de líneas que corresponden a emisiones de sólo algunas longitudes de onda [11]. En la figura 4, se muestra el espectro de emisión del Na (sodio):



Figura 4. Espectro de líneas de emisión del sodio, obtenida de [11]

Espectro con líneas de absorción: El conjunto de líneas espectrales que se obtiene para un elemento concreto es siempre el mismo, incluso si el elemento forma parte de un compuesto complejo, y cada elemento produce su propio espectro diferente al de cualquier otro elemento. Esto significa que cada elemento tiene su propia firma espectral [11]. Si se hace pasar la luz blanca por una sustancia antes de atravesar el prisma sólo pasarán aquellas longitudes de onda que no hayan sido absorbidas por dicha sustancia y se obtendrá su espectro de absorción. Se emplea el espectro de absorción para identificar los elementos químicos que componen la sustancia. En la Figura 5, se muestra el espectro de absorción del sodio:



Figura 5. Espectro de líneas de absorción del sodio, obtenida de [11]

2.2.2 Espectros estelares

Para estudiar la naturaleza de las estrellas, una de las técnicas más utilizadas es la obtención de sus espectros. El espectro estelar como se muestra en la figura 6, contiene información sobre las propiedades físicas de las estrellas, en particular estudiando la intensidad de las líneas de absorción se pueden deducir parámetros estelares como la temperatura y la composición química.

Los espectros de las estrellas son clasificados sobre la base de las intensidades de las líneas espectrales. Las líneas más utilizadas son las de la serie de Balmer, mencionado en [12]. Las líneas del helio neutro y una vez ionizado las líneas del hierro (Fe), líneas de hidrógeno (H), líneas de potasio (K), líneas del doblete de calcio (Ca) ionizado a 396.8 y 293.3 nm. La banda G de la molécula de carbono e hidrógeno (CH) pertenece a la línea de 422.7 nm del calcio (Ca) neutro, y varias líneas de metales alrededor de 422.7 nm y líneas del óxido de titanio (TiO).



Figura 6. Espectro estelar.

2.2.3 Clasificación estelar automática

En términos generales, la clasificación espectral automática se define como la acción ejecutada por un sistema artificial sobre un conjunto de elementos para ordenarlos en clases o categorías.

En relación con lo anterior, para realizar la clasificación de espectros estelares, se adopta el estudio morfológico del espectro visible. Asimismo, en 1864 William Huggins identificó varias de las líneas obtenidas en los espectros de las estrellas con elementos terrestres, demostrando que las estrellas están formadas del mismo material. Desde el comienzo de la espectroscopia, los astrónomos han tratado de crear un sistema para clasificar los espectros estelares, a partir del hecho de que un

gran número de estrellas presentan patrones similares que se pueden identificar en sus espectros [13].

2.2.4 Clasificación espectral de Harvard

El tipo espectral estelar, conocido también como clasificación espectral de Harvard, ya que lo comenzó a esbozar Edward Charles Pickering de la Universidad de Harvard en el año 1890, y que perfeccionó Annie Jump Cannon de la misma universidad en 1901, es la clasificación estelar más utilizada en Astronomía. La clasificación consiste en dividir los espectros estelares que presentan ciertas similitudes en clases espectrales [13]. En 1891, Edward Pickering ordenó los espectros estelares desde la O hasta la M en función de las intensidades de sus líneas de absorción, como se muestran a continuación:

Clase O: Líneas de helio, el oxígeno y el nitrógeno, además de las del hidrógeno, comprende estrellas muy calientes, e incluye tanto las que muestran espectros de línea brillante del hidrógeno y helio como las que muestran líneas más oscuras de los mismos elementos.

Clase B: Líneas de helio que alcanzan la máxima intensidad en la subdivisión B2 y palidecen progresivamente en subdivisiones más altas. La intensidad de las líneas del hidrógeno aumenta de forma constante en todas las subdivisiones. Esta clase está representada por la estrella de Épsilon Orión.

Clase A: Comprende las llamadas estrellas de hidrógeno con espectros dominados por las líneas de absorción del hidrógeno. Una estrella típica de este grupo es Sirio.

Clase F: En este grupo destacan las llamadas líneas de hidrógeno y calcio. Una estrella notable de esta clase es Delta Aquila.

Clase G: Comprende estrellas con fuertes líneas de hidrógeno y calcio, las líneas del hidrógeno son menos prominentes. También están presentes los espectros de muchos metales, en especial el del hierro, el Sol pertenece a esta clase y por ello a las estrellas G se les denomina “estrellas de tipo solar”.

Clase K: Son estrellas que tienen fuertes líneas de calcio y otras que indican la presencia de otros metales. Esta clase está tipificada por la estrella Arturo.

Clase M: Espectros dominados por líneas que indican la presencia de óxidos metálicos, sobre todo las del óxido de titanio. El final violeta del espectro es menos intenso que el de las estrellas K. La estrella Betelgeuse es típica de esta clase.

Dentro de cada clase espectral se puede establecer una secuencia numérica de las subclases, según Meadows en [13]. Cada clase espectral, excepto la clase O se divide en 10 subclases que se designan con cifras desde 0 hasta 9 y que se colocan después de la designación de la clase espectral, por ejemplo, B8, A0, G5. La clase espectral O se divide en subclases desde O4 hasta O9.5. La clasificación espectral permite asignar cada estrella a una clase, grupo o categoría donde se encuentran las estrellas de características similares, con las mismas propiedades químicas, como se muestra en la figura 7.

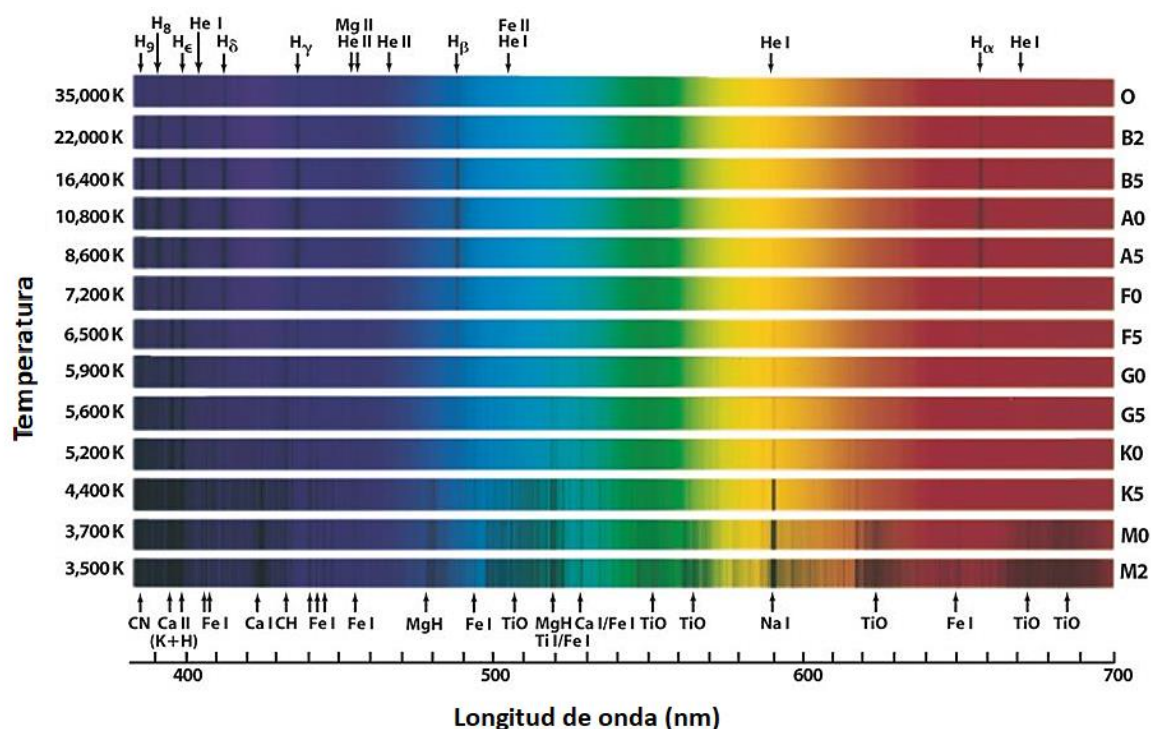


Figura 7. Clasificación espectral de Harvard, obtenida de [13]

2.3 Procesamiento de imágenes digitales

El procesamiento digital de imágenes consiste en operaciones que transforman una imagen en otra donde se resalta cierta información de interés, y se atenúa o elimina información irrelevante para la aplicación. Así, las tareas del procesamiento de imágenes comprenden la supresión de ruido, mejoramiento de contraste, eliminación de efectos no deseados en la captura como iluminaciones o distorsiones por efectos ópticos o de movimiento, mapeos geométricos, transformaciones de color, etcétera [14].

Estas transformaciones varían dependiendo de la imagen y del método o métodos que se apliquen y tendrán repercusión en facilitar o no la etapa tanto de segmentación como de extracción de características de la imagen, como se muestra en la figura 8.

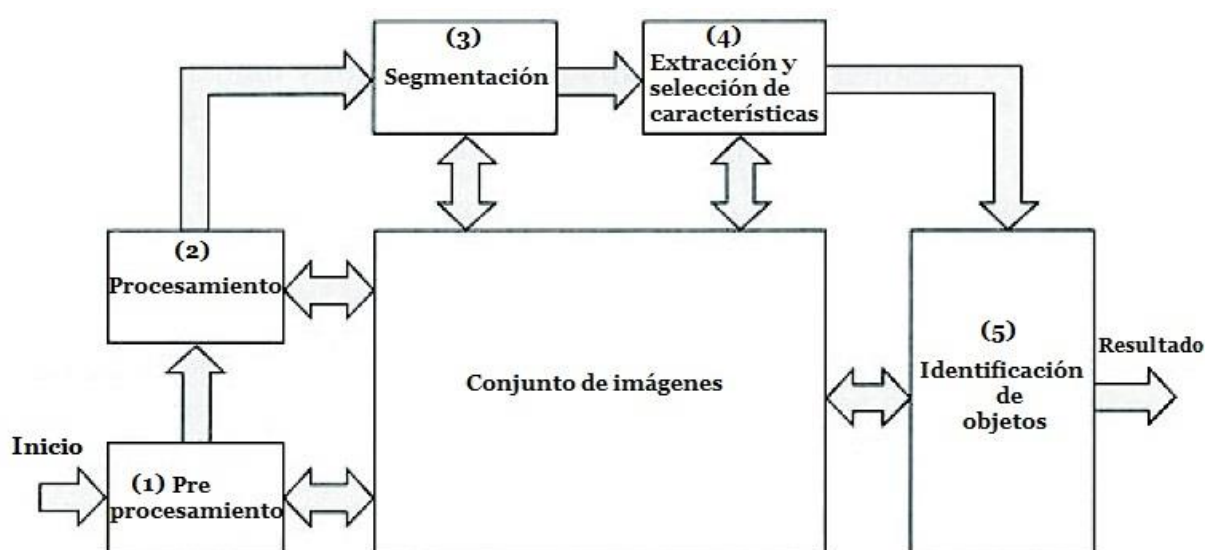


Figura 8. Etapas en el procesamiento de imágenes, obtenida de [15]

2.3.1 Operaciones orientadas al punto

Las operaciones orientadas al punto transforman la imagen modificando un pixel a la vez, en general sin importar el estado de los pixeles vecinos [16]. La transformación se puede aplicar a toda la imagen o a una región de ella. A continuación, se definen los conceptos.

Filtros de aclarado: La carencia de contraste en las imágenes utilizadas en este trabajo puede deberse a una multitud de factores que se desconocen. La idea en las técnicas de aumento de contraste o filtros de aclarado es aumentar el rango dinámico de los niveles de gris en la imagen [14]. Una transformación típica empleada para la mejora del contraste es la representada por el estiramiento del histograma.

Estiramiento del histograma: De acuerdo con Torres en [17], el estiramiento de contraste o estiramiento de histograma se define en la ecuación 1.

$$y = (Max - Min) \frac{x-a}{b-a} + Min \quad (1)$$

Dónde:

y el valor de gris de la imagen resultante.

x el valor de gris de la imagen original.

a valor mínimo de nivel de gris de la imagen original.

b valor máximo de nivel de gris de la imagen original.

Max el valor máximo de nivel de gris que se desea en y .

Min el valor mínimo de nivel de gris que se desea en y .

2.3.2 Operaciones espaciales

Las operaciones espaciales consisten en un procesamiento de la forma [18], de acuerdo a la ecuación 2.

$$v = T(u, u_1, u_2, \dots, u_{p_2-1}) \quad (2)$$

Donde u y v tienen el mismo significado, es decir, valores original y transformado respectivamente de un cierto pixel, y (u_1, \dots, u_{p_2-1}) son los valores de intensidad de los pixeles vecinos del pixel bajo estudio. Una vez más, la función T es común para todos los pixeles de la imagen. Las operaciones espaciales son de tipo convolutivo, esta afirmación es sólo parcialmente cierta; no obstante, cuando sea cierta, la función T será una función lineal de los valores de intensidad de entrada como se muestra en la ecuación (3).

$$v = \lambda_u + \sum_{i=1}^{p_2-1} \lambda_{iu} u_i \quad (3)$$

Como es natural el tipo de operación espacial que se lleve a cabo dependerá de la elección de los coeficientes $(\lambda, \lambda_1, \dots, \lambda_{p_2-1})$ [18]. Por ello, dado que los coeficientes son los que definen enteramente la operación, es práctica habitual resumir estos operadores mediante una máscara de valores, la cual consiste en una rejilla de $P \times P$ pixeles, típicamente con P impar y centrada en torno al pixel bajo estudio, en la que cada casilla contiene el coeficiente por el que se multiplicará a cada pixel que intervenga en la operación [14]. De esta forma, con sólo una inspección de la máscara se puede interpretar rápidamente en qué consiste el operador en cuestión [16]. No obstante, existirán casos en los que las operaciones espaciales no sean lineales.

Realce de bordes: Si bien es cierto que bajo esta denominación pueden entrar todos los operadores que lleven a cabo operaciones de filtrado paso alto, este nombre se reserva para operadores de tipo gradiente, es decir, operadores de primera derivada, y en general, con carácter direccional. En la práctica, cuando se trabaja con imágenes digitalizadas, no se lleva a cabo la operación en un dominio continuo, sino que se aproxima a una versión discreta. Tal versión discreta se obtiene mediante el uso de máscaras [16]. Algunas de las más conocidas y por ende las más empleadas, son el operador de Sobel que técnicamente es un operador diferencial discreto que calcula una aproximación al gradiente de la función de intensidad de una imagen, calcula el gradiente de la intensidad de una imagen en cada punto (píxel). Así, para cada punto, este operador da la magnitud del mayor cambio posible, la dirección de éste y el sentido desde oscuro a claro. Enfatiza la contribución de los píxeles más cercanos al píxel bajo estudio, el resultado muestra cuán probable es que represente un borde en la imagen y, también, la orientación a la que tiende ese borde [16].

2.3.3 Operaciones basadas en regiones

Las operaciones orientadas a la región transforman la imagen modificando un píxel a la vez y toman en cuenta para dicha transformación los píxeles vecinos [16]. Y como es natural la transformación se puede aplicar a toda la imagen o a una región de ella.

Recorte de la imagen: El corte o extracción de objetos en imágenes digitales es un paso esencial para métodos posteriores, según Ortiz en [16]. Una vez llegado a este proceso, se tienen objetos candidatos dada su área. A su vez se liberará carga de procesamiento computacional al disminuir la cantidad de píxeles y reducir el tamaño de la imagen.

2.3.4 Operaciones geométricas

Una operación geométrica es aquella que cambia el tamaño, forma u orientación de una imagen [16]. En general no se puede considerar un filtro, pero corresponden a transformaciones útiles en el procesamiento digital de imágenes. Por ejemplo, cuando se requiere escalar, orientar, o empalmar una imagen se pueden aprovechar dichas operaciones.

Escalar: Opera sobre la imagen entera, agranda o reduce el tamaño físico de la imagen cambiando el número de píxeles que contiene.

Rotación: Consiste girar la imagen en un ángulo definido, el uso de esta transformación se encuentra por ejemplo en la alineación de las imágenes respecto a cierta referencia, para realizar una presentación adecuada en la solución de algunos problemas [16].

2.4 Extractor de características WND-CHARM

Los autores Singh y Carpenter en [19], explican que un extractor de características, es un algoritmo de clasificación en el que se extrae una variedad de características morfológicas obtenidas directamente en imágenes completas, evitando la necesidad de segmentación. Dos ventajas que presentan son: su capacidad para capturar una amplia variedad de aspectos morfológicos de la imagen y la ausencia de requisitos para la segmentación. Se ha demostrado que tiene un buen desempeño en una amplia gama de problemas de clasificación, tales como imágenes biológicas mencionado por Lior [20], reconocimiento de letras por Orlov [21], clasificación facial por Samaria [22], clasificación de galaxias por morfología de Shamir [23] y clasificación de espectros estelares por Avila [24]. Esto proporciona por primera vez una evaluación realista y confiable de toda la estrategia de clasificación de imágenes.

WND-CHARM es un extractor de características y clasificador de funciones de imagen multiuso que se puede aplicar a una amplia variedad de tareas de clasificación de imágenes sin modificaciones ni ajustes [25]. Los tipos de parámetros calculados por este método se dividen en cuatro categorías: descomposiciones polinómicas, características de alto contraste, estadísticas de píxeles y texturas. El primer paso en la clasificación de imágenes es representar el contenido de la imagen como un conjunto de valores numéricos (características). Debido a la amplia gama de tareas realizadas por los clasificadores generalizados de imagen, el número de funciones calculadas durante el entrenamiento es mucho mayor que en los clasificadores específicos de la tarea. En la descomposición polinomial, se genera un polinomio que aproxima la imagen a cierta fidelidad, y los coeficientes de este polinomio se utilizan como descriptores del contenido de la imagen. Las características de textura informan sobre la variación de intensidad entre píxeles de varias direcciones y resoluciones. Características de alto contraste, como bordes y objetos, comprenden estadísticas sobre el número de objetos, la distribución espacial, el tamaño, la forma, etcétera. Las estadísticas de píxeles se basan en la distribución de las intensidades de píxeles dentro de la imagen e incluyen histogramas y momentos. El vector de características comprende, en su formato reducido 1025 descriptores y en su versión extendida 2920 descriptores, cada uno informa sobre un aspecto diferente del contenido de la imagen. Todas las funciones se basan en imágenes en escala de grises, por lo que la información de color no se utiliza actualmente. No se ha intentado normalizar este espacio variable, muchas de las características pueden ser interdependientes y no pueden considerarse ortogonales. Las estadísticas de objetos se calculan solo en la imagen original, mientras que los polinomios de Zernike se calculan en la imagen original con la transformada rápida de Fourier (FFT), pero no en otras transformaciones. Las principales características que se calculan con WND-CHARM se muestran en la figura 9.

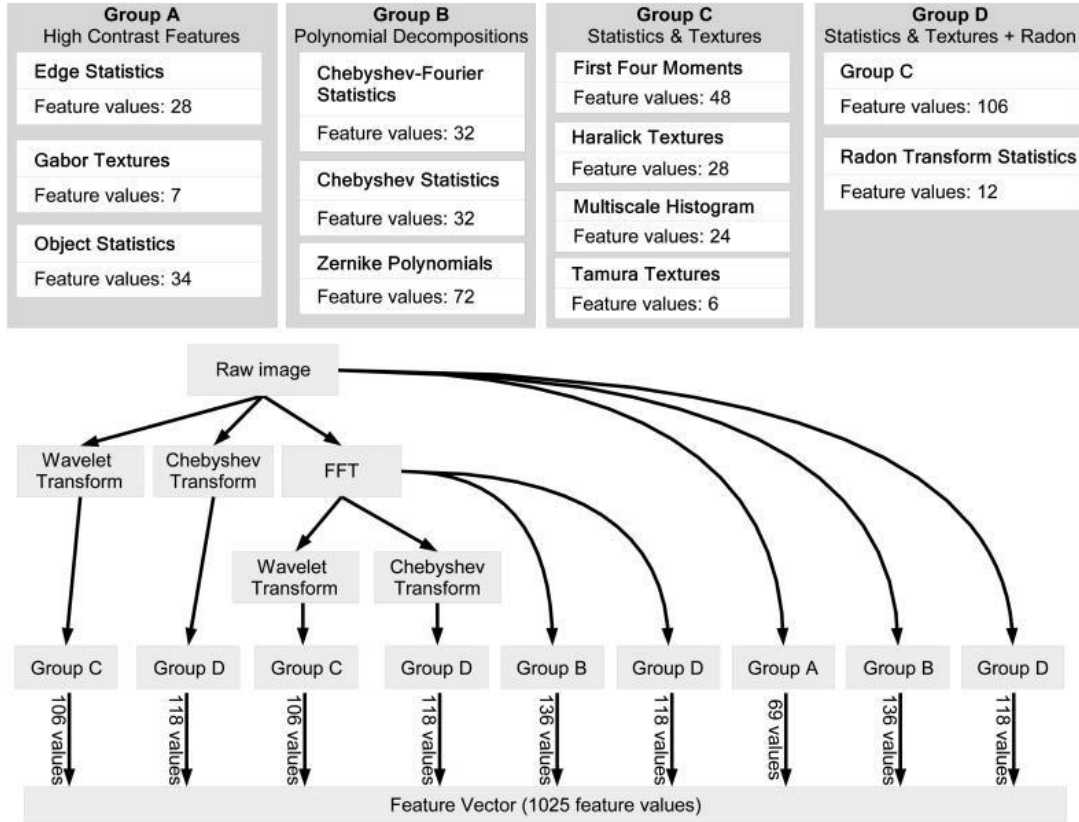


Figura 9. Características calculadas por el proceso WND-CHARM, obtenida de [25]

2.4.1 Características de la imagen basadas en los coeficientes de Chebyshev

De acuerdo con Boyd, los descriptores de Chebyshev son características relacionadas con la frecuencia y la continuidad de toda la imagen que se pueden aplicar en diferentes dominios (imagen en bruto, transformada de Fourier y Wavelet) basados en polinomios de Chebyshev [26].

Los polinomios de Chebyshev según la ecuación (3) son ampliamente utilizados para fines de aproximación según Gradshtein en [27].

$$T_n(x) = \cos(n \cdot \arccos(x)) \quad (3)$$

Dónde:

$T_n(x)$ es un polinomio de grado n evaluado en el punto x

Para cualquier función dada, de acuerdo con Gradshtein en [28], se puede generar su aproximación Chebyshev como se muestra en la ecuación (4).

$$f(x) \cong \sum_{n=0}^N \alpha_n T_n(x) \quad (4)$$

Dónde:

T_n es un polinomio de grado n

α es el coeficiente de expansión

Dado que estos polinomios son ortogonales según Gradshtein; los coeficientes de expansión α_n se pueden expresar a través del producto escalar, como se muestra en la ecuación (5) [28].

$$\alpha_n = \{f(x), T_n(x)\} \quad (5)$$

Dónde:

α es el coeficiente de expansión

α_n es un producto escalar de dos polinomios

Para una imagen dada I , su aproximación bidimensional a través de los polinomios de Chebyshev, se expresa en la ecuación (6).

$$I_{i,j} = I(x_i, y_i) \cong \sum_{n,m=0}^N \alpha_{nm} T_n(x_i) T_m(y_i) \quad (6)$$

Dónde:

I es la imagen

i, j son las filas y columnas (imagen)

α_{nm} es un producto escalar de dos polinomios

T_m polinomio de grado m

T_n polinomio de grado n

El algoritmo toma dos transformaciones, primero para las filas de la imagen, luego para las columnas de la matriz resultante, así como para la implementación de la transformada rápida de Fourier (FFT 2D) [28]. Los coeficientes de Chebyshev son utilizados por el extractor WND-CHARM, como una transformación (con órdenes que coincidan con las dimensiones de la imagen) y como conjunto de estadísticas. $N = 20$, por lo que el coeficiente tiene dimensiones (1×400) . Las características de la imagen son el histograma de 32 bandejas de los 400 coeficientes. Dado que las características de Chebyshev se compilan en la imagen sin procesar y en la imagen transformada de Fourier, el número total de descriptores de imagen agregados al vector de características es 64.

2.4.2 Características de la imagen basadas en la transformada de Chebyshev-Fourier

La transformada 2D de Chebyshev-Fourier se define en coordenadas polares y utiliza dos tipos diferentes de transformadas ortogonales para sus dos variables: distancia y ángulo, según Boyd en [29]. La distancia se aproxima con los polinomios de Chebyshev, y el ángulo se deduce mediante armónicos de Fourier, como se describe en la ecuación (7). Los armónicos de Fourier se utilizan para descomponer una función, señal u onda periódica como suma infinita o finita de funciones de señales u ondas armónicas o sinusoidales; es decir, es una serie trigonométrica.

$$\Omega_{n,m}(r, \varphi) = T_n \left(\frac{2r}{R} - 1 \right) \exp^{im\varphi}, 0 \leq r \leq R \quad (7)$$

Dónde:

$\Omega_{n,m}$ armónicos de Fourier

r distancia

φ ángulo

T_n polinomio de grado n

n orden del armónico ($n > 1$)

R resonancia

im imagen

Para la imagen dada I la transformada se describe en la ecuación (8).

$$I_{ij} \rightarrow I(r_k, \varphi l) \cong \sum_{m=-\frac{N}{2}}^{\frac{N}{2}} \sum_{n=0}^N \beta_{nm} \Omega_{nm}(r_k, \varphi l) \quad (8)$$

Dónde:

I imagen

β_{nm} coeficientes de Chebyshev

Ω_{nm} coeficientes de Fourier

2.4.3 Características de la imagen basadas en momentos de Comb.

Los momentos de Comb en imágenes 2D, según Sheng en [30]. A_{nm} del orden n con repetición m se definen usando coordenadas polares (r, θ) dentro del círculo unitario en una imagen como se describe en la ecuación (9).

$$A_{nm} = \int_0^{2\pi} \int_0^1 C_n(r) \exp(-jm\theta) f(r, \theta) r dr d\theta \quad (9)$$

Dónde:

A_{nm} del orden n con repetición m

$C_n(r)$ valores reales de los momentos de Comb

2.5 Selección de características algoritmo Relief-F

El algoritmo fue desarrollado en 1992 por Kira y Rendell [31], adopta un enfoque de método de filtro para la selección de características, que es notablemente sensible a estas interacciones. Originalmente fue diseñado para su aplicación en problemas de clasificación binaria con características discretas o numéricas.

Relief-F calcula una puntuación para cada función dada, se puede aplicar para clasificar y seleccionar las funciones de puntuación más altas. Alternativamente, estas puntuaciones pueden aplicarse como ponderaciones de características para guiar el modelado posterior. La puntuación de la característica central, se basa en la identificación de las diferencias de valor entre el vecino más cercano y pares de instancias. Si se observa una diferencia del valor en un par de instancias vecinas con la misma clase, es un acierto, la puntuación de la característica disminuye. Alternativamente, si se observa una diferencia del valor en un par de instancias vecinas con diferentes valores de la clase, es un fallo, y la puntuación de la característica aumenta.

De acuerdo con Kira, el algoritmo propuesto Relief-F hace un proceso iterativo asignando ponderaciones a cada característica y se puede obtener una aproximación de valores de una imagen completa utilizando la ecuación (10). Donde $C_{n,m}$ corresponden a los coeficientes de Chebyshev calculados por el n y m grados de la imagen. T_n y T_m son los polinomios correspondientes calculados para cada fila i y columna j de la imagen. Finalmente, se obtiene el valor promedio dividido por el resultado del ancho w multiplicado por la altura de la imagen h en píxeles.

$$C_{n,m} = \frac{\sum_{i,j} I_{i,j} T_n(i) T_m(j)}{w * h} \quad (10)$$

Dónde:

$C_{n,m}$ = Coeficientes de Chebyshev calculando n y m grados de la imagen

i, j = filas y columnas (imagen)

I = imagen

T_n = polinomio n para cada columna i de la imagen

T_m = polinomio m para cada fila j de la imagen

w = ancho de la imagen

h = altura de la imagen

2.6 Técnicas de aprendizaje automático

Machine Learning (aprendizaje automático) es una técnica de análisis de datos que enseña a los ordenadores a hacer lo que resulta natural para las personas y los animales: aprender de la experiencia [32]. Los algoritmos de aprendizaje automático emplean métodos de cálculo para “aprender” información directamente de los datos sin depender de una ecuación predeterminada como modelo. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje.

En [32], el aprendizaje automático emplea dos tipos de técnicas: el aprendizaje supervisado, que entrena un modelo con datos de entrada y salida conocidos para que pueda predecir salidas futuras y el aprendizaje no supervisado, que encuentra patrones ocultos o estructuras intrínsecas en los datos de entrada, como se muestra en la figura 10.

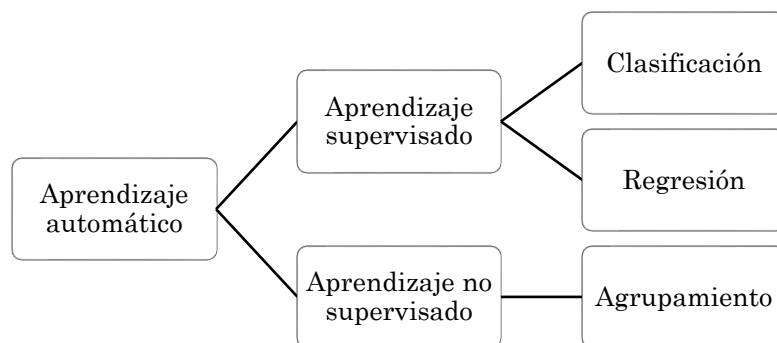


Figura 10. Técnicas de aprendizaje automático: se incluyen el aprendizaje supervisado y el aprendizaje no supervisado

En la tabla 1, se presenta un resumen de las técnicas más utilizadas en clasificación digital de imágenes.

Técnicas de aprendizaje automático	Características	Ejemplos
Basadas en píxeles	Cada píxel se considera puro, es decir se considera una única firma espectral.	No supervisados: K-means, clustering. Supervisados: Maximum likelihood, Minimum distance-to-means, Mahalanobis distance, Parallelepiped, K- nearest neighbors.
Basadas en sub-píxeles	Cada píxel es considerado como una mezcla de firmas espectrales.	Fuzzy classification, neural networks, regression modeling, regression tree analysis, spectral mixture analysis.

Tabla 1. Técnicas de aprendizaje automático, obtenido de [10]

2.6.1 Máquina de soporte vectorial

Una máquina de soporte vectorial (SVM) construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre dos clases de datos se amplía al máximo [33]. Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión.

Las máquinas de soporte vectorial pertenecen a una clase de algoritmos de aprendizaje automático denominados métodos kernel y también se conocen como máquinas kernel. El entrenamiento de una máquina de soporte vectorial consta de dos fases:

- 1) Transformar los predictores (datos de entrada) en un espacio de características altamente dimensional. En esta fase es suficiente con especificar el kernel; los datos nunca se transforman explícitamente al espacio de características. Este proceso se conoce comúnmente como el truco kernel.
- 2) Resolver un problema de optimización cuadrática que se ajuste a un hiperplano óptimo para clasificar las características transformadas en dos clases. El número de características transformadas está determinado por el número de vectores de soporte.

Para construir la superficie de decisión solo se requieren los vectores de soporte seleccionados de los datos de entrenamiento. Una vez entrenados, el resto de los datos de entrenamiento son irrelevantes [33].

2.6.2 Árboles de decisión

Es uno de los enfoques de modelado predictivo que se utiliza en estadística, minería de datos y aprendizaje automático. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de

condiciones que ocurren de forma sucesiva, para la resolución de un problema. Los modelos de árbol donde la variable destino puede tomar un conjunto finito de valores se denominan árboles de clasificación. En estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan las conjunciones de características que conducen a esas etiquetas de clase. Los árboles de decisión, donde la variable destino puede tomar valores continuos (por lo general números reales) se llaman árboles de regresión, mencionado en [32]. De forma más concreta, se puede decir que los árboles de decisión son diagramas de decisiones secuenciales que muestran sus posibles resultados. Estos ayudan a determinar cuáles son sus opciones al mostrar distintas decisiones y resultados.

2.6.3 Vecino más cercano

En el reconocimiento de patrones, el algoritmo de vecinos k –más cercanos (k -NN) es un método no paramétrico utilizado para la clasificación y regresión [34]. En ambos casos, la entrada consta de los k ejemplos de entrenamiento más cercanos en el espacio de características. La salida depende de si k -NN se utiliza para la clasificación o regresión: En la clasificación k -NN, la salida es una membresía de clase; un objeto se clasifica por mayoría de votos de sus vecinos, y el objeto se asigna a la clase más común entre sus k vecinos más cercanos (k es un número entero positivo, típicamente pequeño). Si $k = 1$, entonces el objeto simplemente se asigna a la clase de ese vecino más cercano. En la regresión k -NN, la salida es el valor de propiedad del objeto. Este valor es el promedio de los valores de sus k vecinos más cercanos. Tanto para la clasificación como para la regresión, se puede usar una técnica útil para asignar peso a las contribuciones de los vecinos, de modo que los vecinos más cercanos contribuyan más al promedio que los más distantes. Por ejemplo, un esquema de ponderación común consiste en dar a cada vecino un peso de $1/d$, donde d es la distancia al vecino [34].

Los vecinos se toman de un conjunto de objetos para los que se conoce la clase (para la clasificación k -NN) o el valor de la propiedad del objeto (para la regresión k -NN). Esto puede considerarse como el conjunto de entrenamiento

para el algoritmo, aunque no se requiere ningún paso de entrenamiento explícito. Una peculiaridad del algoritmo k -NN es que es sensible a la estructura local de los datos.

2.6.4 Análisis discriminante lineal

El análisis discriminante lineal busca combinaciones lineales de variables que expliquen mejor los datos, intenta explícitamente modelar la diferencia entre las clases de datos. Funciona cuando las mediciones realizadas en variables independientes para cada observación son cantidades continuas [35].

El análisis discriminante se usa cuando los grupos se conocen a priori. Cada caso debe tener una puntuación en una o más medidas de predicción cuantitativa, y una puntuación en una medida de grupo. En términos simples, el análisis de función discriminante es una clasificación: el acto de distribuir cosas en grupos, clases o categorías del mismo tipo.

2.6.5 Redes Neuronales

Este tipo de clasificación se basa en el uso de redes neuronales artificiales, una neurona artificial es un objeto lógico (software) que recibe diversas entradas, realiza una suma ponderada de las mismas y produce una salida a partir de la aplicación de una función umbral a la medida ponderada [36]. La conexión de las salidas de unas neuronas con la entrada de otras, se le conoce como una red neuronal.

En las redes neuronales la información fluye en una única dirección, de la capa de entrada, pasando por la o las capas ocultas hacia la capa de salida. Cada neurona de cualquier capa está conectada (en la dirección de flujo) con todas las neuronas de la siguiente capa. Un ejemplo típico de una red neuronal es *Back Propagation Neural Network (BPNN)* se muestra en la figura 11.

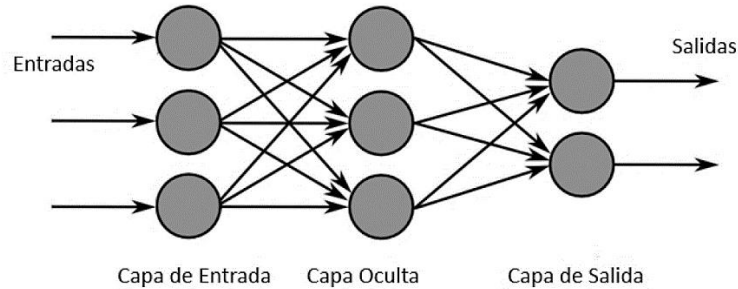


Figura 11. Red Neuronal *Back Propagation*, obtenida de [8]

Castillo, menciona en [10], que el proceso de clasificación sigue tres fases:

- 1) **Entrenamiento:** Se introducen píxeles cuya clase se conoce, y se compara la salida con la realidad. Con el resultado de la comparación se modifican los coeficientes de ponderación de todas las neuronas para obtener la respuesta adecuada, es decir valores de 1 en la clase correcta y cero en la clase incorrecta.
- 2) **Estabilización:** Al inicio del entrenamiento los factores de ponderación cambian rápidamente, pero conforme este se desarrolla los factores comienzan a estabilizarse, en el momento en que no se modifican más, finaliza la fase de entrenamiento.
- 3) **Clasificación:** Se introducen píxeles cuya clase no se conoce y se adjudican a la clase con la respuesta más alta.

2.7 Técnicas de evaluación

Una forma eficaz de evaluar los resultados de las técnicas de aprendizaje automático, es utilizando técnicas de evaluación, que miden la eficiencia del aprendizaje, es decir, intentar extraer alguna medida que informe de lo bien o mal que se está clasificando. La exactitud se medirá con la ecuación (11).

$$exactitud = \frac{\#aciertos}{\#objetosClasificados} \quad (11)$$

2.7.1 Matriz de confusión

En el campo de la inteligencia artificial una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en una técnica de aprendizaje automático. Cada columna de la matriz representa el número de predicciones de cada clase y muestra la exactitud alcanzada, mientras que cada fila representa a las instancias en la clase real, mostrando la precisión obtenida, como se muestra en la figura 12, mencionado en [37]. Uno de los beneficios de las matrices de confusión es que facilitan saber si el sistema está confundiendo dos clases.

	C_1	C_2	C_r
C_1	n_{11}	n_{12}	n_{1r}
C_2	n_{21}	n_{22}	n_{2r}
.				
.				
C_r	n_{r1}	n_{r2}	n_{rr}

Figura 12. Matriz de Confusión, obtenida de [37]

2.7.2 Validación cruzada de K iteraciones

En la validación cruzada de K iteraciones o *K-fold cross-validation* los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos [38]. Lo más común es utilizar la validación cruzada de 10 iteraciones (*10-fold cross-validation*).

Capítulo 3

Trabajos Relacionados

3.1 Introducción del capítulo

En el presente capítulo se detallan los trabajos relacionados que sustentan y forman parte del desarrollo de este proyecto de tesis. Se realiza una revisión de trabajos relacionados en el área de Observación de la Tierra con firmas espectrales. Posteriormente, se detallan trabajos de clasificación de espectros estelares y finalmente de extracción de características. En el primer subtema, se estudian los artículos basados en la obtención de la firma espectral, su importancia y datos relevantes. Por consiguiente, se analizan los trabajos previos de clasificación estelar, las técnicas utilizadas y los resultados obtenidos. Finalmente, se menciona un algoritmo de extracción de características utilizado en la clasificación de galaxias por morfología. En la tabla 2, se muestra una comparativa de herramientas y metodologías relacionadas con este trabajo de tesis; las viñetas que se aprecian en la tabla identifican los métodos y herramientas por nombre del autor.

Autor y año	Análisis y procesamiento de imagen	Identificación y extracción de espectros en imágenes digitales	Técnicas de aprendizaje automático	Catálogo de Pulkovo (señales de espectros estelares)	Metodología de firmas espectrales	WND-CHARM extractor de características	Algoritmo de selección de característica
(Shamir, 2009)			✓			✓	✓
(Peregrina, 2016)			✓	✓			
(Castilla, 2016)	✓		✓	✓	✓		
(Peregrina, 2018)			✓	✓			✓
(Avila, 2018)	✓	✓	✓		✓	✓	✓

Tabla 2. Tabla comparativa.

3.2 Firmas espectrales

Carrillo Castilla define en [10], que una de las aplicaciones más importantes de las firmas espectrales es la clasificación, gracias a las características análogas que posee un objeto es más sencillo categorizarlo en clases. Para su realización existen métodos que requieren de conocimientos previos del área de estudio, y otros donde no es necesario, al mismo tiempo para realizar la clasificación se cuenta con métodos que emplean inteligencia artificial. El resultado final es una imagen digital clasificada que permite la identificación de clases dependiendo del objeto de estudio, se propone un sistema de información geográfica basado en redes neuronales, especializado en procesos correctivos y de valor agregado a sistemas ópticos multiespectrales, que tiene por objetivo desarrollar servicios de preprocesamiento para productores y consumidores de imágenes digitales que requieran de recursos de cómputo suficientes en sus organizaciones. Los resultados del trabajo son: mapas temáticos para su análisis en riesgos y desastres naturales; mediante la implementación de dos colecciones de firmas espectrales, la primera se trata de *Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)* de la NASA, que incluye datos de tres laboratorios el *Johns Hopkins University (JHU)*, *Jet Propulsion Laboratory (JPL)* y *United States Geological Survey (USGS)*, la segunda es una base de datos nacional (BDN) generada por el Instituto Potosino de Investigación Científica y Tecnológica. Como resultados obtenidos del autor Carrillo Castilla, se logró desarrollar un modelo computacional integrado de diversos módulos para el tratamiento de imágenes ópticas multiespectrales, con gran aplicabilidad en el área de geociencias (ciencias de la tierra), se realizó un nuevo algoritmo clasificador basado en redes neuronales y acoplamiento de bases digitales, los resultados preliminares mostraron un alto índice de exactitud del 92%, clasificando en su totalidad firmas espectrales del flujo de calor en la zona del volcán Popocatepetl.

La aplicación de firmas espectrales, según el autor Carrillo Castilla es reportada para una amplia gama de áreas, principalmente Geociencias y Astrofísica, con el fin de alcanzar altos índices de clasificación.

El autor Corrales Andino en [39], menciona la importancia de la obtención de firmas espectrales para el estudio de objetos de la tierra, explica que hay tres tipos de firmas espectrales (monobandas, multibandas y de origen opcional) en función a las características de captura de la información por parte del sensor que se utiliza; en este trabajo se utilizó el Sistema de Información Geográfica ArcGis, versión 9.2 con licencia. Se obtuvo un conjunto de firmas espectrales de cubiertas de suelo de la ciudad universitaria de Tegucigalpa (UNAH) de Honduras, como apoyo para la identificación, clasificación de coberturas y uso del suelo. El objetivo del trabajo propuesto era establecer un método de captura de firmas espectrales en un ambiente natural. Los inconvenientes surgieron en la exportación del conjunto de datos de firmas espectrales, ya que los archivos fueron guardados en extensiones propias del sistema utilizado y no se podían exportar a ningún otro sistema.

3.3 Espectros estelares

La clasificación estelar es un tema importante en tareas astronómicas, como el estudio de poblaciones estelares. Actualmente, existen métodos que involucran diferentes áreas, como el procesamiento de imágenes, extracción de características, aprendizaje automático (*machine learning*) y técnicas de clasificación (árboles de decisión, redes neuronales, máquinas de soporte vectorial, algoritmos genéticos, entre otros). En el presente trabajo, “*Automatic Stellar Spectral Classification with Multiple Intelligent Classifiers*” en [4], los autores proponen una clasificación de objetos estelares basada en tres estrategias principales: perceptrones multicapa (MLP), vecino más cercano (K-NN) y máquinas de soporte vectorial (SVM). Se propuso un criterio de clasificación que evalúa los resultados de las técnicas de representación dispersa y determina la clasificación final de los espectros. Esta metodología demuestra su capacidad para alcanzar niveles de clasificación comparables con otras metodologías automáticas como son: el Coeficiente de Correlación Máxima (MCC) y Redes Neuronales Artificiales (ANN). La principal conclusión de este trabajo son los resultados de precisión para los enfoques SVM con tasas de 88.1% utilizando un kernel lineal. Por otro lado, se estudiaron las

técnicas de representación dispersa y el aprendizaje de diccionario en “*Automatic stellar Spectral Classification Via Sparse Representations and Dictionary Learning*” [5], los autores proponen mejorar la precisión total de cinco clases de clasificación espectral estelar. Estos se aplican utilizando el algoritmo OMP (*Orthogonal Matching Pursuit*) para encontrar una solución aproximada y la mejor representación de clases para una señal de entrada Y . Por lo tanto, la clasificación dispersa se basa en el reconocimiento de las características comunes de un tipo estelar en particular a través de la construcción de una base de señales entrenadas, se reportan 529 muestras de espectros estelares para las clases [B, A, F, G, K] basadas en el sistema de Harvard [13], tales muestras pertenecen al Catálogo Espectrofotométrico Pulkovo, mencionado en [40]. Este trabajo concluye, que los mejores resultados están directamente relacionados con una buena representación de los datos por cada espectro estelar, que es representado dada una señal Y de entrada, y se utiliza tanto para el entrenamiento como para la prueba de la metodología propuesta. El uso de un conjunto de datos de señales es utilizado en la mayoría de los enfoques de clasificación de espectros estelares [4]. Dentro de este contexto, se menciona que el Catálogo de Pulkovo es un concentrado de una colección de datos, de estrellas observadas en 20 años en diferentes partes del mundo, tales como Chile, Armenia, Tadjikistan y Bolivia [40]. Los datos del Catálogo se encuentran en unidades de flujo de energía absoluta y se pueden usar para determinar algunos parámetros físicos de las estrellas, como las temperaturas efectivas y los diámetros angulares. En el Catálogo de Pulkovo, se utiliza una amplia gama de estrellas de diferentes tipos espectrales, donde se pueden seleccionar estrellas de cada clase para derivar su distribución de energía media inherente a una subclase dada [40]. Para estrellas peculiares, se puede investigar el efecto de las líneas de absorción, su composición química y los efectos de enrojecimiento interestelar y circunstancial.

3.4 Extracción de características

Por otra parte, existen diferentes enfoques que se implementan para otro tipo de tareas en Astronomía, por ejemplo la clasificación automática de galaxias en imágenes según su morfología, que fue adoptada con éxito por el proyecto Galaxy Zoo [23]. Muestra un enfoque de clasificación automática, donde las imágenes de galaxias son adquiridas por Sloan Digital Sky Survey (SDSS) [41], en el proyecto de Galaxy Zoo se visualizan las imágenes en formato JPEG.

Los autores Shamir y Lior, proponen una etapa de extracción de características, utilizando WND-CHARM para cada una de las imágenes de galaxias. Este extractor calcula valores numéricos relacionados con las características más importantes de la imagen, que distinguen entre tipos morfológicos de galaxias. Posteriormente, los valores se usan para diferentes enfoques de aprendizaje automático y alcanzan altas tasas de precisión. Las características encontradas en las imágenes del proyecto de Galaxy Zoo en [23], incluyen características de alto contraste (estadísticas de objeto, estadísticas de borde, Filtros Gabor) entre otros. Debido a estos resultados, se propone utilizar un extractor de características en este trabajo de tesis. En *WND-CHARM—An Open Source Utility for Biological Image Analysis*, se menciona que es una herramienta de código abierto para el análisis de imágenes y el software funciona extrayendo primero los descriptores del contenido de la imagen, las transformaciones de imagen y las transformadas de imagen compuesta. En seguida, se seleccionan las características más importantes, formando finalmente un vector de características de cada imagen y se usa para la clasificación de objetos, así como también para la medición de similitud [20].

3.5 Conclusión del capítulo

En conclusión, en este capítulo se realizó una revisión del trabajo relacionado referente a firmas espectrales, el autor Carrillo Castilla en [10], expone la colección de firmas espectrales que utilizó para la clasificación del flujo de calor en la zona del volcán Popocatepetl, posteriormente menciona que la aplicación de firmas espectrales es utilizada para una amplia gama de áreas, tales como la Geociencia y Astrofísica entre otras disciplinas, con el fin de alcanzar altos índices de clasificación.

En trabajos previos, de acuerdo con el autor G. A Alekseeva en [40], existen catálogos de espectroscopia que contienen parámetros específicos de espectros estelares, estos datos forman un vector de señales con el objetivo de clasificarlos. Un aspecto a considerar por los autores Cruz Vega y Díaz Hernández, fue clasificar cinco de las siete clases posibles de espectros estelares con un conjunto de señales, obtenido del catálogo de Pulkovo. Para mejorar los resultados obtenidos, los autores concluyen que se requiere de un mayor número de muestras de espectros estelares para clasificar las siete clases posibles, utilizando máquinas de soporte vectorial (SVM), una técnica de aprendizaje supervisado, como se menciona en [4] y [5].

La ventaja que se propone en esta tesis; es extraer las características propias de la imagen, obteniendo la firma espectral de un espectro estelar, se utilizó la colección de placas astronómicas tomadas por la cámara Schmidt de Tonantzintla, esta colección posee un acervo de imágenes digitalizadas que contienen cúmulos de estrellas.

Capítulo 4

Trabajo desarrollado

4.1 Introducción del capítulo

Como se ha descrito anteriormente, la clasificación estelar de una región del cielo es un proceso que requiere mucho tiempo debido a la gran cantidad de objetos estelares que se concentran en una sola imagen, y se requiere dedicación de tiempo completo por parte de un experto. Por lo tanto, con el uso de técnicas automáticas se acelera este proceso sin necesidad de intervención humana, de una manera rápida y eficiente.

En el capítulo anterior, los autores Carrillo y Hernández mencionan la importancia de realizar una clasificación con la firma espectral de un objeto, obtenida de una imagen digital, mencionan los inconvenientes de compatibilidad al realizar su trabajo en un sistema de información geográfico con licencia. En Carrillo, se explica que con la implementación de un sistema de información geográfico, gratuito y abierto al público se tendría una herramienta de fácil acceso; la desventaja es establecer un formato general en el cual se pueda guardar la información y sea compatible con otros sistemas de información. Los autores Cruz y Hernández, proponen un conjunto de datos obtenido del Catálogo Espectrofotométrico Pulkovo, los datos contenidos en este catálogo forman parte de diferentes observaciones en tres países del mundo.

Con la finalidad de clasificar las imágenes de espectros estelares, presentes en las placas astronómicas; en este capítulo se propone un conjunto de datos de firmas espectrales para su posterior clasificación. Para cumplir con los objetivos planteados en un principio en este trabajo de tesis; en primera instancia, se analiza el acervo de imágenes astronómicas, se definen las operaciones realizadas para el análisis y procesamiento de imágenes digitales. Posteriormente, se detalla el

proceso de extracción y selección de características que se llevó a cabo en las imágenes utilizadas, y finalmente se describen los sistemas de aprendizaje automático que se utilizaron para la clasificación de espectros estelares.

4.2 Conjunto de datos propuesto

Los conjuntos de datos que se presentan en trabajos previos para la clasificación de espectros estelares, contienen información de observatorios de diferentes partes del mundo. En esta tesis se propone trabajar con un conjunto de datos de firmas espectrales, extraídas del acervo de imágenes astronómicas que se encuentran al resguardo del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Como parte del trabajo realizado, se utilizaron veinte imágenes astronómicas. En la tabla 3, se muestra el conjunto de imágenes de espectros estelares que se extrajeron de las imágenes de las placas astronómicas, siendo la primera fila la clase a la que pertenece el espectro estelar, y la segunda fila el número total de sub imágenes de espectros estelares.

Clases	O	B	A	F	G	K	M	Total
Muestras	33	960	784	567	675	499	49	3563

Tabla 3. Colección de imágenes de espectros estelares por tipo espectral.

De acuerdo a la clasificación del sistema de Harvard en [2], del conjunto que se mencionó anteriormente, se tienen cinco clases [B, A, F, G, K] de las cuales se clasificaron en cuatro subclases respectivamente, teniendo un total de 20 subclases, como se muestra en la tabla 4.

Clases	B				A				F				G				K			
Subclases	1	3	4	5	0	3	5	6	1	3	5	7	0	3	6	9	1	2	4	5
Muestras	51	49	49	48	39	48	51	50	49	51	54	46	49	50	55	51	52	55	54	56

Tabla 4. Colección de imágenes de espectros estelares por subtipo espectral

En la tabla 5, se muestra el conjunto de imágenes astronómicas utilizadas para la clasificación de espectros estelares, se muestra el nombre de la placa y la cantidad de espectros estelares extraídos de cada imagen. Las imágenes están guardadas en formato TIFF, y contienen desde veinte hasta trecientas sub imágenes de espectros estelares.

Número	Nombre de la placa	Espectros extraídos
1	AC1058	141
2	AC1598	101
3	AC1310	192
4	AC3192	116
5	AC1599	50
6	AC1315	190
7	AC1338	187
8	AC1571	357
9	AC3883	147
10	AC1336	195
11	AC2656	227
12	AC1353	128
13	AC3275	141
14	AC1060	185
15	AC1565	125
16	AC1343	225
17	AC1572	230
18	AC1579	243
19	AC1350	181
20	AC3865	231

Tabla 5. Conjunto de imágenes astronómicas.

En la figura 13, se enlistan los pasos que se siguieron para poder obtener el conjunto de datos para la etapa de clasificación.

- 1) Teniendo como primer paso un ejemplo de una imagen astronómica de la cual se extrae e identifica la imagen de un espectro estelar; para identificar los espectros de la imagen astronómica se realizó un proceso visual de etiquetado que recopila varias sub imágenes de la imagen original.
- 2) En el paso dos, se puede observar la sub imagen espectro estelar resultante de 50 x 300 píxeles, esta imagen contiene información química del espectro y muestra sus líneas de absorción.
- 3) Por último, en el tercer paso se obtiene el conjunto de datos de firmas espectrales.

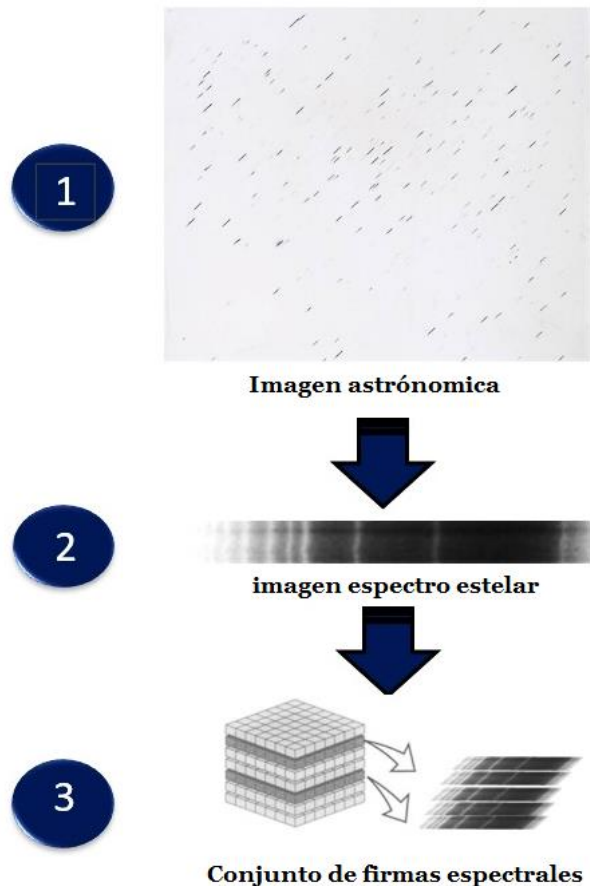


Figura 13. Obtención del conjunto de datos propuesto

4.3 Acervo de imágenes astronómicas

Para este trabajo de tesis, se utilizaron imágenes de placas de tipo espectral, como se muestra en la figura 14, es una imagen digitalizada de muy alta resolución con dimensiones de 12124 x 12062 píxeles con una profundidad de 48 bits, esta imagen forma parte del acervo de imágenes astronómicas que posee el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Actualmente se cuenta con el acervo de 14,613 placas astronómicas que fueron digitalizadas en el año 2000, mencionado por Díaz Hernández en [42].



Figura 14. Placa tipo espectral.

4.4 Formato utilizado en el almacenamiento de las imágenes

Las imágenes Tagged Image File Format (TIFF por sus siglas en inglés) descritas en [43], es el formato de imagen que se utiliza en este trabajo debido a las siguientes características:

- La mínima pérdida de calidad de la imagen, lo hace ideal para tratar las imágenes astronómicas debido a que se requiere el mayor detalle posible para el reconocimiento de los espectros, mencionado en [44].
- La pérdida de datos que se tiene al momento de abrir, guardar o modificar un archivo de imagen se ve disminuida gracias a la profundidad en la que se almacenan los datos.
- Es el formato predeterminado en el que se encuentra digitalizada la imagen.

4.5 Representación de las imágenes utilizadas en escala de grises

Cuando la función de imagen mide los valores de intensidad luminosa, se tienen imágenes en escala de grises y se pueden representar con un valor numérico de esta magnitud por cada pixel. Aunque no es estrictamente necesario, es costumbre utilizar números naturales para la intensidad lumínica o brillo [45]. Los motivos fundamentales de utilizar imágenes en escala de grises son dos:

- 1) El número finito y no muy elevado de niveles producidos por los captadores físicos de imágenes.
- 2) La considerable economía de almacenamiento de estos números frente a los reales.

Además, y con el objetivo de adaptar lo mejor posible la escala de almacenamiento en disco duro, el número de niveles suele ser una potencia de 2, siendo la más frecuente de $2^8 = 256$ niveles.

4.6 Procesamiento de imágenes

Con base en el Marco Teórico, el procesamiento de imágenes aplicado en las imágenes astronómicas comprende el mejoramiento de contraste y el realce de bordes. En algunos casos se necesitó mejorar la nitidez de las imágenes, en la tabla 6, se presentan ejemplos de imágenes del acervo de placas astronómicas; en la imagen (c) se aprecian con claridad los espectros estelares, y se busca obtener las demás imágenes (a) y (b) análogas a esta.


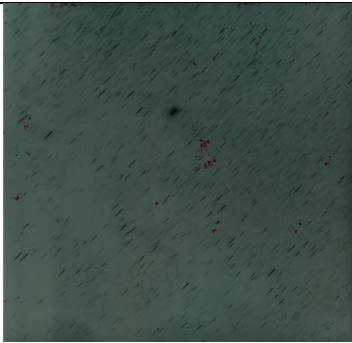

Imagen (a)	Imagen (b)	Imagen (c)
		

Tabla 6. Placas digitalizadas.

4.6.1 Mejoramiento de contraste

En la tabla 7, se muestran ejemplos de imágenes astronómicas, con el fin de aumentar el rango dinámico de los niveles de gris en la imagen, se les aplicó un filtro de aclarado, posteriormente se convirtió a escala de grises. Una transformación típica es la representada por el estiramiento del histograma, el objetivo al emplear este filtro es aminorar el ruido presente en las imágenes astronómicas; se muestra la imagen original (a), y el resultado obtenido con el filtro se observa en la imagen (b), al realizar una comparación entre ambas imágenes (a) y (b), se puede observar que la imagen (b) mejoró notablemente.



Imagen original (a)	Estiramiento del histograma (b)
	

Tabla 7. Mejoramiento de contraste

4.6.2 Realce de bordes

En la tabla 8, se muestra un ejemplo del resultado del operador de Sobel, aplicado a una imagen digitalizada, se puede observar con detenimiento que las líneas que aparecen en la imagen (b) de forma diagonal, se muestran más pronunciadas con respecto a la imagen original (a).



Imagen original (a)	Realce de borde (b)
	

Tabla 8. Operador de Sobel.

4.7 Segmentación de espectros

Para la segmentación de espectros estelares, se transforma la imagen modificando un pixel a la vez y se toma en cuenta para dicha transformación los pixeles vecinos; la transformación se puede aplicar a toda la imagen, pero en este caso es a una región de ella. El corte o extracción de los objetos, es un paso esencial para que en métodos posteriores se adquiera solamente la imagen del espectro estelar. Como primer paso, se carga la imagen completa, y posteriormente se realiza el recorte dependiendo de la región en donde ésta se ubique; crear un recorte ayuda a delimitar el área sobre la cual se desea trabajar, a su vez libera carga de procesamiento computacional al disminuir la cantidad de pixeles [10]. El recorte se realizó de manera semiautomática, como se muestra en la figura 15.

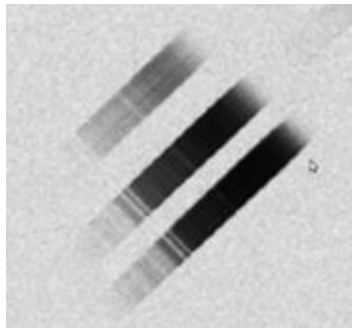


Figura 15. Recorte de espectros estelares

4.7.1 Rotación de la imagen

Una vez obtenidos los espectros estelares de forma individual y haber implementado la extracción de los mismos se debe rotar la imagen a 43 grados, esto para poder visualizar la imagen del espectro de manera horizontal. En la figura 16, se muestra la imagen obtenida.



Figura 16. Imagen rotada de espectros estelares

Finalmente, se repite el paso anterior para recortar la imagen y obtener como resultado una imagen de 300 x 50 pixeles, como se muestra en la figura 17. Esta imagen es el resultado final que forma parte del conjunto de datos de firmas espectrales que se propone en este trabajo de tesis.



Figura 17. Imagen de un espectro estelar.

4.8 Extracción de características

Utilizando el extractor de características WND-CHARM propuesto en [25]. Se logró obtener un vector de 1x2920 características, como se muestra en la figura 18, siendo los coeficientes de Chebyshev, la transformada de Chebyshev-Fourier y los momentos de Comb por morfología, los mejores descriptores extraídos de diferentes dominios de la imagen, para este caso de clasificación.

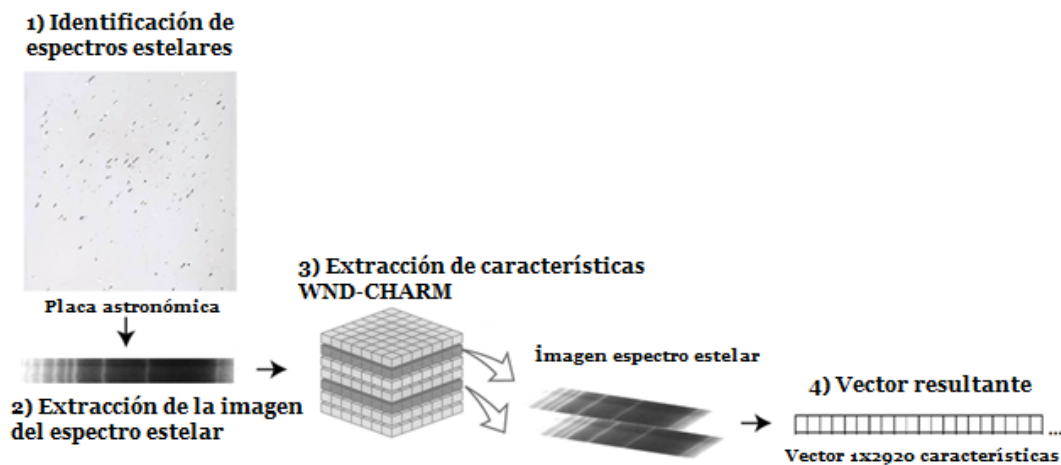


Figura 18. Proceso de extracción de características.

4.8.1 Descriptores obtenidos

En la imagen 19, se utilizan los descriptores de imagen de los momentos de Comb, que son valores relacionados con la forma de los objetos en las imágenes. En este sentido, se produce un vector con 32 valores; el proceso se repite utilizando las imágenes de espectros estelares, con la transformación de Fourier. De este modo, el orden de transformación máximo no supera $n, m = 20$, por lo tanto, se calculó un total de $(20 \times 20) = 400$ coeficientes de Chebyshev por imagen.

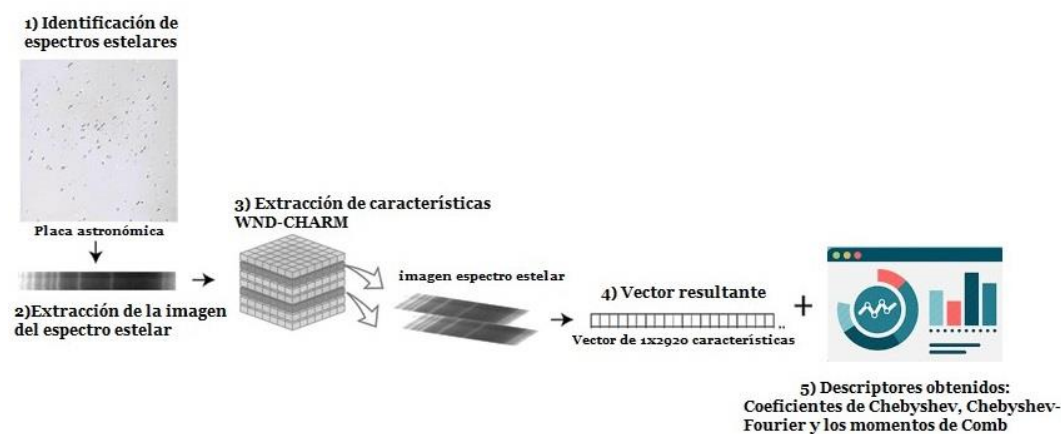


Figura 19. Descriptores obtenidos de WND-CHARM

4.9 Selección de características

En el proceso anterior de extracción de características, se utilizó el extractor WND-CHARM, obteniendo como resultado un vector de 1x2920 características, conforme al vector obtenido se decide utilizar el algoritmo Relief-F de selección de características, para reducir las dimensiones del vector debido a su alto procesamiento de cómputo. En la figura 20, se indica conforme a la numeración los pasos a seguir. Al utilizar el algoritmo Relief- F, se obtuvo un vector de 1x25 dimensiones con las características más importantes de cada imagen de espectros estelares y de acuerdo con el proceso de extracción de características, los descriptores de los coeficientes de Chebyshev, la transformada de Chebyshev-Fourier y los momentos de Comb son útiles para las siete clases del método de clasificación que se presenta.

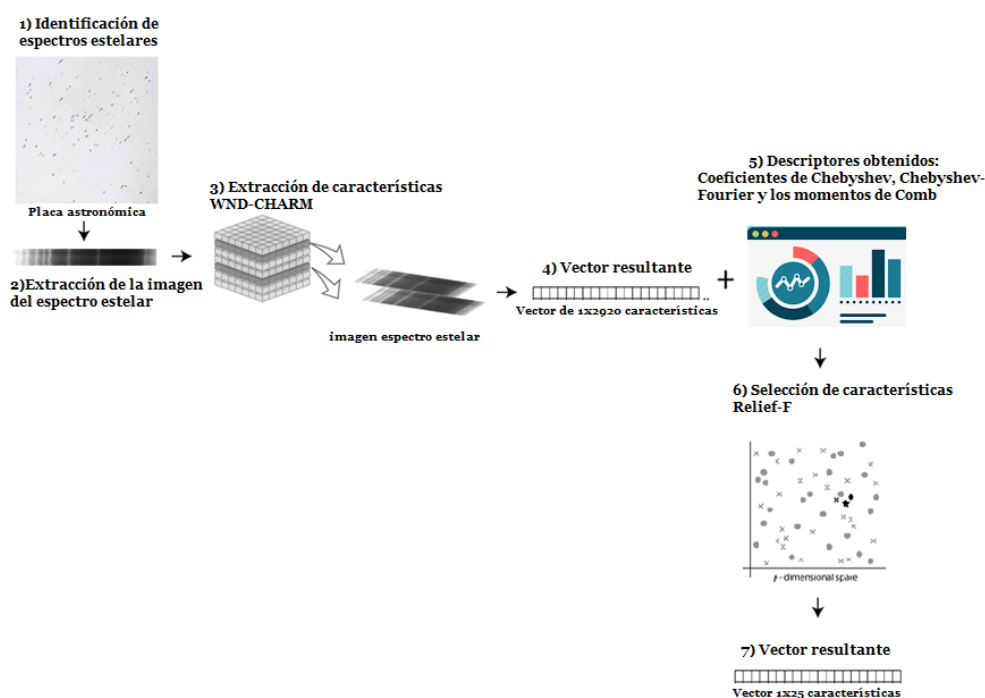


Figura 20. Proceso de selección de características

4.10 Técnicas de aprendizaje automático utilizadas

En la tabla 9, se muestran las características de las cinco técnicas utilizadas de aprendizaje automático.

Técnica	Velocidad de predicción	Uso de memoria	Interpretabilidad
Árboles de decisión 	Rápido	Pequeña	Fácil
Análisis discriminante 	Rápido	Pequeño para lineal, grande para cuadrático	Fácil
Máquinas de soporte vectorial 	Medio para lineales. Lento para los demás.	Medio para lineales. Todos los demás: medio para multiclase, grande para binario	Fácil para SVM lineal. Difícil para todos los demás tipos de kernel.
El vecino más cercanos 	Lento para cúbico. Medio para los demás.	Medio	Difícil
Redes neuronales 	Rápido	Medio	Fácil

Tabla 9. Técnicas de aprendizaje automático

4.11 Clasificación de espectros estelares

La clasificación de imágenes digitales representa la fase final de este método, en otras palabras, tiene como objetivo asignarles a los píxeles valores que sirvan como identificadores de la clase donde se deben incluir.

Para la clasificación automática de espectros estelares, se propone que los descriptores principales de estas imágenes, se basen principalmente en los coeficientes de la transformada de Chebyshev-Fourier, y los valores absolutos de los coeficientes complejos que se utilizan para la descripción de la imagen.

El propósito de los descriptores obtenidos, es capturar componentes de baja frecuencia (áreas a gran escala con transiciones suaves de intensidad) del contenido de la imagen. El orden más alto de polinomio utilizado es $N = 23$, y el coeficiente ahora se reduce al agrupar a 1×25 longitudes. Dado que las características de Chebyshev-Fourier se compilan en la imagen en bruto y en la imagen transformada de Fourier, WND-CHARM proporciona estos descriptores de imagen para el vector de características [29].

Posteriormente, se realiza la clasificación del conjunto de datos obtenido de firmas espectrales, se agrupan en varias clases según su composición química y temperatura.

Para finalizar con lo propuesto, se muestra en la figura 21 el proceso completo para la clasificación de espectros estelares, indicando por números consecutivos el orden de este método.

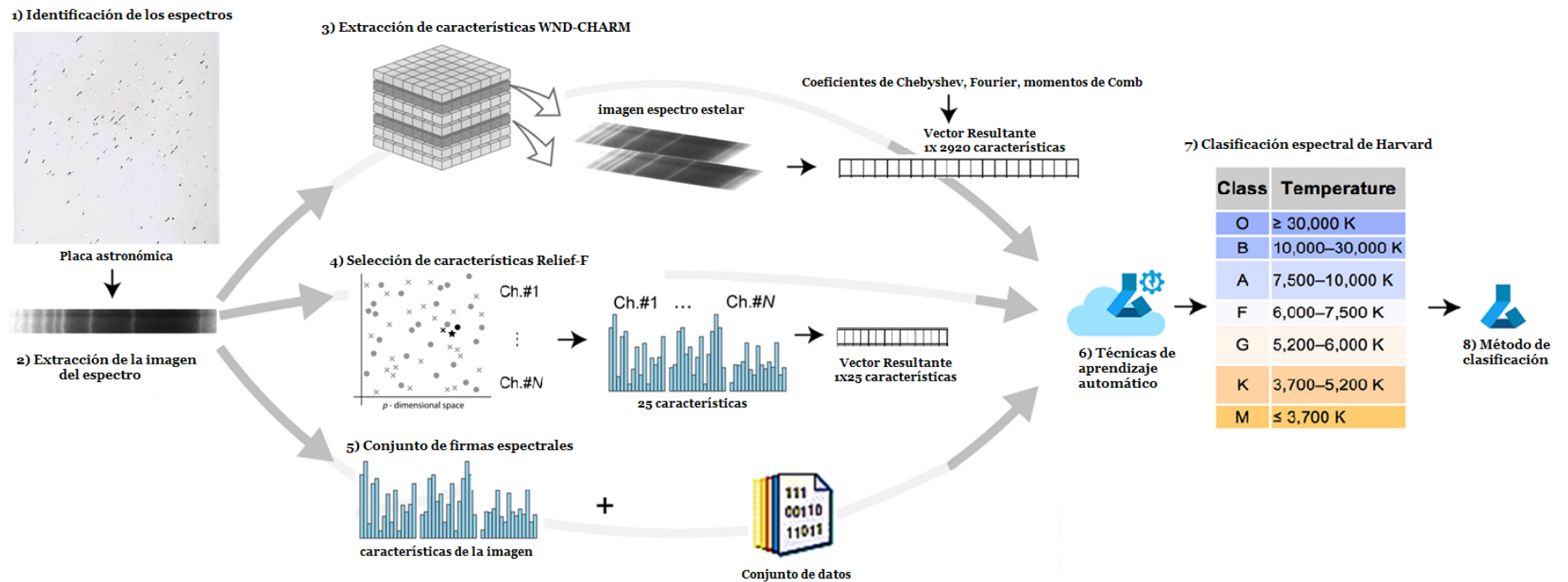


Figura 21. Método de clasificación de espectros estelares

4.12 Conclusión del capítulo

Como producto del trabajo realizado en este capítulo y debido a la necesidad de realizar una clasificación automática de espectros estelares, se propuso un conjunto de datos de firmas espectrales, extraído del acervo de imágenes digitalizadas que se encuentra al resguardo del INAOE, a diferencia de otros conjuntos de datos reportados.

Se describen las técnicas de aprendizaje automático para la clasificación de espectros estelares con base al Sistema de Harvard, con la finalidad de identificar de forma automática la clase a la que pertenecen los espectros.

Estas técnicas permiten la obtención de firmas espectrales extraídas de imágenes, y el análisis que se realiza es con mayor cantidad de información en el menor tiempo posible.

En el siguiente capítulo, se presenta la clasificación realizada de este método con cinco técnicas de aprendizaje.

Capítulo 5

Experimentos y Resultados

Con el fin de reconocer y clasificar múltiples objetos de espectros estelares, se utilizaron técnicas de aprendizaje automático, como son: árboles de decisión (*decision trees*), el vecino más cercano (*K-NN k-nearest neighbors*), la máquina de soporte vectorial (*SVM support vector machines*), el análisis discriminante (*discriminant analysis*) y redes neuronales (*ANN neural networks*). Se utilizaron estas técnicas para comparar la exactitud y precisión de los resultados. Los experimentos se realizaron con validación cruzada en 10 partes (*10-fold cross-validation*), para optimizar los parámetros del modelo, y se ajuste a los datos de entrenamiento tan bien como pueda.

5.1 Clasificación de espectros estelares

La clasificación de este trabajo se basó en el aprendizaje automático (*machine learning*), y de acuerdo con el extractor de características WND-CHARM se obtuvo un vector de 1x2920 características, siendo los coeficientes de Chebyshev, la transformada de Chebyshev-Fourier y los momentos de Comb por morfología, los mejores descriptores extraídos de diferentes dominios de la imagen, para este caso de clasificación. En la selección de Relief-F, se obtuvieron las 25 características más importantes de cada imagen, logrando mejorar la exactitud y precisión en la mayoría de los casos. Los experimentos se dividieron en 4 fases, se enlistan a continuación:

- 1) Clasificación de cinco grupos de clases de tipos espectrales
- 2) Clasificación con los coeficientes de Chebyshev y los momentos de Comb
- 3) Clasificación de siete grupos de clases de tipos espectrales
- 4) Subclasificación de espectros estelares

5.1.1 Fase I. Clasificación de cinco grupos de clases de tipos espectrales

Para esta primera fase, se fusionaron las clases espectrales de tipo O y M con las clases de tipo B y K respectivamente, conforme a la relación de temperatura y composición química. En la tabla 3, del capítulo 4 se presentó el conjunto de muestras que se obtuvo por cada clase espectral, teniendo 33 muestras en la clase de tipo O y 960 de la clase de tipo B, al fusionar las clases quedó un total de 993 muestras, de igual forma para la clase de tipo M, se tienen 49 muestras y al fusionarla con la clase de tipo K, se tiene un total de 548 muestras, como se muestra en la tabla 10.

Clases	OB	A	F	G	KM	Total
muestras	993	784	567	675	548	3563

Tabla 10. Grupo con cinco clases de espectros estelares.

Las matrices de confusión para las pruebas de clasificación de 5 clases que usan las 25 características más importantes se muestran en las tablas [11, 12, 13, 14, 15] con las cinco técnicas utilizadas. En cada técnica se mejoraron los resultados de las clases y estos fueron consistentes en las pruebas. Las matrices de confusión se leen de izquierda a derecha, la fila indica la clase verdadera, la columna indica la salida del clasificador. Ejemplo: cada entrada esta en negrita, y da el número de aciertos que clasificó correctamente, al final de cada fila se muestra el porcentaje de precisión que se alcanzó. Como resultado, todas las correctas clasificaciones se muestran en la diagonal parte superior izquierda a inferior derecha, teniendo como resultado la exactitud de dicho clasificador. Todo lo que esta fuera de esa diagonal es una clasificación incorrecta de algún tipo de clase. Estos fueron los primeros resultados en un problema de clasificación de cinco clases.

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	646	200	68	58	21	65.06%
A	166	492	55	49	21	62.84%
F	157	89	214	77	30	37.74%
G	92	72	52	372	86	55.19%
KM	81	54	51	266	94	17.22%
Exactitud						51.02%

Tabla 11. Matriz de Confusión Árbol de Decisión (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	990	0	0	3	0	99.70%
A	0	783	0	0	0	100.00%
F	3	3	561	0	0	98.94%
G	0	0	9	665	0	98.66%
KM	0	0	0	336	210	38.46%
Exactitud						90.06%

Tabla 12. Matriz de Confusión KNN (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	990	0	3	0	0	99.70%
A	6	777	0	0	0	99.23%
F	6	0	561	0	0	98.94%
G	9	0	0	409	256	60.68%
KM	0	0	0	324	222	40.66%
Exactitud						83.05%

Tabla 13. Matriz de Confusión SVM (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	429	192	232	102	38	43.20%
A	142	442	118	55	26	56.45%
F	102	75	279	74	37	49.21%
G	88	59	131	286	110	42.43%
KM	60	36	94	235	121	22.16%
Exactitud						43.70%

Tabla 14. Matriz de Confusión de Análisis Discriminante (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	846	69	48	12	18	85.20%
A	75	627	39	18	24	80.08%
F	69	60	396	18	24	69.84%
G	27	24	29	411	183	60.98%
KM	21	15	24	222	264	48.35%
Exactitud						71.40%

Tabla 15. Matriz de Confusión de RN (5 clases)

Con esto se concluye que, de las técnicas de aprendizaje propuestas, el mejor resultado se logró utilizando el vecino más cercano (K-NN) con una tasa de exactitud del 90.06% clasificando la mayoría de espectros, seguido de la máquina de soporte vectorial (SVM) con un 83.05%. Los resultados muestran una mejora significativa en las primeras cuatro clases OB, A, F y G.

5.1.2 Fase II. Clasificación con los coeficientes de Chebyshev y los momentos de Comb

Se repitieron los experimentos con cinco clases y se formó un vector con las 25 características más importantes para cada imagen, como se muestra en la tabla 16.

Núm.	Coeficientes	Características de la imagen
1	415	1 Chebyshev-Fourier Coefficients (Fourier ()) [31]
2	1230	1 Comb Moments (Chebyshev (Fourier ())) [6]
3	1227	17 Comb Moments (Chebyshev (Fourier ())) [3]
4	1238	10 Comb Moments (Chebyshev (Fourier ())) [14]
5	1145	1 Haralick Textures (Chebyshev ()) [17]
6	806	17 Comb Moments (Wavelet ()) [6]
7	241	3 Comb Moments () [1]
8	263	7 Comb Moments () [23]
9	1242	3 Comb Moments (Chebyshev (Fourier ())) [18]
10	257	1 Comb Moments () [17]
11	2780	2 Comb Moments (Wavelet (Edge ())) [4]
12	1237	0 Comb Moments (Chebyshev (Fourier ())) [13]
13	1541	0 Chebyshev-Fourier Coefficients (Fourier (Wavelet ())) [29]
14	1228	2 Comb Moments (Chebyshev (Fourier ())) [4]
15	242	7 Comb Moments () [2]
16	1524	0 Chebyshev-Fourier Coefficients (Fourier (Wavelet ())) [12]
17	1231	18 Comb Moments (Chebyshev (Fourier ())) [7]
18	2386	0 Chebyshev-Fourier Coefficients (Fourier (Edge ())) [26]
19	807	0 Comb Moments (Wavelet ()) [7]
20	414	0 Chebyshev-Fourier Coefficients (Fourier ()) [30]
21	1567	0 Chebyshev Coefficients (Fourier (Wavelet ())) [23]
22	2240	6 Comb Moments (Edge ()) [24]
23	1098	3 Comb Moments (Chebyshev ()) [18]
24	261	3 Comb Moments () [21]
25	240	10 Comb Moments () [0]

Tabla 16. Las 25 características más importantes obtenidas de WND-CHARM

En las tablas [17, 18, 19, 20, 21] se muestran los resultados de clasificación en matrices de confusión para esta segunda fase, donde se utilizaron los coeficientes de Chebyshev y los momentos de Comb como características vectoriales por imagen.

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	716	129	46	80	22	72.10%
A	141	556	43	23	20	71.01%
F	109	79	309	61	9	54.50%
G	73	49	29	431	92	63.95%
KM	43	44	22	316	121	22.16%
Exactitud						59.87%

Tabla 17. Matriz de Confusión Árbol de Decisión (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	987	3	0	3	0	99.40%
A	0	777	3	3	0	99.23%
F	3	0	564	0	0	99.47%
G	0	0	0	668	6	99.11%
KM	0	0	0	330	216	39.56%
Exactitud						90.15%

Tabla 18. Matriz de Confusión KNN (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	990	0	3	0	0	99.70%
A	9	774	0	0	0	98.85%
F	6	0	561	0	0	98.94%
G	3	0	0	417	254	61.87%
KM	0	0	0	329	217	39.74%
Exactitud						83.05%

Tabla 19. Matriz de Confusión SVM (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	990	0	3	0	0	99.70%
A	13	770	0	0	0	98.34%
F	44	4	519	0	0	91.53%
G	15	0	0	349	310	51.78%
KM	2	0	0	273	271	49.63%
Exactitud						81.36%

Tabla 20. Matriz de Confusión Análisis Discriminante (5 clases)

Tipo Espectral	OB	A	F	G	KM	Precisión
OB	996	3	9	9	6	97.36%
A	9	771	0	3	0	98.47%
F	12	6	540	3	6	95.24%
G	6	3	3	509	153	75.52%
KM	0	3	3	195	345	63.19%
Exactitud						87.98%

Tabla 21. Matriz de Confusión Redes Neuronales (5 clases)

Para la segunda fase de clasificación se repitieron las cinco clases [OB, A, F, G, KM], utilizando los coeficientes de Chebyshev y los momentos de Comb por morfología como características vectoriales por imagen, de las técnicas de aprendizaje utilizadas, se concluye que la técnica que obtuvo los mejores resultados en esta fase fue nuevamente el vecino más cercano (K-NN) con una tasa de exactitud del 90.15% superando el resultado de la fase anterior y clasificando de manera acertada la mayoría de espectros, seguido de las redes neuronales con una tasa de exactitud del 87.98%. Los resultados muestran una mejora significativa en las primeras cuatro clases OB, A, F y G.

5.1.3 Fase III. Clasificación de siete grupos de clases de tipos espectrales

Para la última fase se clasificaron las 7 clases [O, B, A, F, G, K, M] por tipo espectral, en la tabla 22 se presentan el número de muestras. En esta prueba se utilizaron los coeficientes de Chebyshev y los momentos de Comb por morfología como características vectoriales por imagen, empleando las mismas técnicas de aprendizaje automático.

Clases	O	B	A	F	G	K	M	Total
muestras	33	960	784	567	675	499	49	3563

Tabla 22. Colección de imágenes de espectros estelares por tipo espectral.

En las tablas [23, 24, 25, 26, 27] se muestran los resultados en matrices de confusión de esta prueba.

Tipo Espectral	O	B	A	F	G	K	M	Precisión
O	0	14	5	7	3	4	0	0.00%
B	1	64	136	56	63	20	0	18.82%
A	0	157	549	28	29	20	0	70.11%
F	1	93	95	282	73	21	2	49.74%
G	0	53	33	41	483	63	1	71.66%
K	1	40	37	22	304	94	0	18.88%
M	1	4	2	3	29	2	7	14.58%
Exactitud								50.25%

Tabla 23. Matriz de Confusión Árbol de Decisión (7 clases)

Tipo Espectral	O	B	A	F	G	K	M	Precisión
O	33	0	0	0	0	0	0	100.00%
B	0	957	0	3	0	0	0	99.69%
A	0	3	777	3	0	0	0	99.23%
F	0	3	0	564	0	0	0	99.47%
G	0	0	0	0	674	0	0	100.00%
K	0	3	0	0	330	165	0	33.13%
M	0	0	0	0	0	0	48	100.00%
Exactitud								90.32%

Tabla 24. Matriz de Confusión KNN (7 clases)

Tipo Espectral	O	B	A	F	G	K	M	Precisión
O	29	4	0	0	0	0	0	87.88%
B	0	958	0	2	0	0	0	99.79%
A	0	12	771	0	0	0	0	98.47%
F	0	3	0	564	0	0	0	99.47%
G	0	3	0	0	403	268	0	59.79%
K	0	6	0	0	328	164	0	32.93%
M	0	10	0	0	0	0	38	79.17%
Exactitud								82.15%

Tabla 25. Matriz de Confusión SVM (7 clases)

Tipo Espectral	O	B	A	F	G	K	M	Precisión
O	24	6	0	3	0	0	0	72.73%
B	4	705	121	74	38	18	0	73.44%
A	0	96	624	35	16	12	0	79.69%
F	1	122	41	334	55	14	0	58.91%
G	0	34	17	46	440	137	0	65.28%
K	3	25	8	15	262	182	3	36.55%
M	0	5	0	3	14	1	25	52.08%
Exactitud								65.51%

Tabla 26. Matriz de Confusión Análisis Discriminante (7 clases)

Tipo Espectral	O	B	A	F	G	K	M	Precisión
O	30	3	0	0	0	0	0	90.91%
B	0	951	3	6	0	0	0	99.06%
A	0	3	777	0	0	3	0	99.23%
F	0	6	0	561	0	0	0	98.94%
G	0	6	3	9	506	150	0	75.07%
K	0	6	0	3	183	306	0	61.45%
M	0	0	0	0	0	3	45	93.75%
Exactitud								89.14%

Tabla 27. Matriz de Confusión Redes Neuronales (7 clases)

Se concluye que en esta fase los mejores resultados fueron obtenidos con la técnica del vecino más cercano (K-NN) con una tasa de exactitud del 90.32% superando el resultado de la fase anterior y clasificando de manera acertada la mayoría de espectros, seguido de la técnica de redes neuronales con una tasa de exactitud del 89.14%. Los resultados muestran una mejora significativa en las siete clases O, B, A, F, G, K y M.

5.2 Clasificación de subclases de espectros estelares

Los astrónomos dividieron los tipos de estrellas con las letras O, B, A, F, G, K y M, para nombrarlas de acuerdo a sus componentes químicos y temperatura. Las subclases del 0 al 9 se utilizan para indicar sucesiones en el modelo de cada clase. Cada clase espectral, excepto la clase O se divide en 10 subclases que se designan con cifras desde 0 hasta 9 y se colocan después de la designación de la clase espectral; por ejemplo: B8, A0, G5; excepto la clase espectral O, que se divide en subclases desde O4 hasta O9.5 [2]. La sub-clasificación espectral permite asignar cada espectro estelar a una clase donde se agrupan con características similares, y poseen las mismas propiedades químicas, o bien la misma firma espectral.

5.2.1 Fase IV. Clasificación de subclases de espectros estelares

Para la sub clasificación de espectros estelares, se repite el proceso de la clasificación principal; utilizando los coeficientes de Chebyshev y los momentos de Comb como características vectoriales. Se muestran en la tabla 28, las clases B, A, F, G y K con sus respectivas subclases.

Clases	B				A				F				G				K			
Subclases	1	3	4	5	0	3	5	6	1	3	5	7	0	3	6	9	1	2	4	5
muestras	51	49	49	48	39	48	51	50	49	51	54	46	49	50	55	51	52	55	54	56

Tabla 28. Colección de imágenes de espectros estelares por subtipo espectral

En la tabla 29, en la primera columna se muestran las clases estelares por tipo espectral, en la segunda columna la clasificación de subclases, finalmente en la tercera columna se muestran los resultados obtenidos, en este caso en particular, solo se muestra la exactitud obtenida por cada técnica que se utilizó.

Clasificación de subclases de espectros estelares		
Árbol de decisión		
Clase	Subclase	Exactitud
B	1, 3, 4, 5	63.24%
A	0, 3, 5, 6	64.93%
F	1, 3, 4, 7	62.43%
G	0, 3, 6, 9	63.86%
K	1, 2 ,4, 5	60.01%
El vecino más cercano		
Clase	Subclase	Exactitud
B	1, 3, 4, 5	90.12%
A	0, 3, 5, 6	90.23%
F	1, 3, 4, 7	89.13%
G	0, 3, 6, 9	90.24%
K	1, 2 ,4, 5	90.15%
Máquinas de soporte vectorial (SVM)		
Clase	Subclase	Exactitud
B	1, 3, 4, 5	75.78%
A	0, 3, 5, 6	74.17%
F	1, 3, 4, 7	72.21%
G	0, 3, 6, 9	73.86%
K	1, 2 ,4, 5	72.16%
Análisis discriminante		
Clase	Subclase	Exactitud
B	1, 3, 4, 5	70.12%
A	0, 3, 5, 6	71.30%
F	1, 3, 4, 7	73.97%
G	0, 3, 6, 9	71.03%
K	1, 2 ,4, 5	72.34%
Redes Neuronales		
Clase	Subclase	Exactitud
B	1, 3, 4, 5	84.34%
A	0, 3, 5, 6	80.23%
F	1, 3, 4, 7	82.32%
G	0, 3, 6, 9	81.39%
K	1, 2 ,4, 5	86.19%

Tabla 29. Subclasificación de espectros estelares

Se concluye en esta fase, que los mejores resultados obtenidos fueron con la técnica del vecino más cercano (K-NN) con tasas de exactitud del 90.12%, 90.23%, 89.13%, 90.24%, 90.15% superando los resultados de las demás técnicas utilizadas, clasificando de manera acertada la mayoría de espectros.

5.3 Discusión de los resultados obtenidos

El trabajo experimental de la tesis, se dividió en cuatro fases. Se muestran las conclusiones de los resultados obtenidos para cada fase, y se indica en las tablas [26, 27, 28, 29] la exactitud alcanzada de cada técnica.

5.3.1 Fase I

En esta primera fase, se utilizó el extractor WND-CHARM obteniendo un vector de 1x2920 valores. Posteriormente, se utilizó el algoritmo Relief-F adquiriendo las 25 características más importantes de la imagen de un espectro estelar. Esta fase consistió en unir las clases [OB] y [KM] de acuerdo a su temperatura, y fue posible alcanzar un índice de precisión promedio del 67.84% de las técnicas utilizadas, los resultados se muestran en la tabla 30.

Técnica utilizada	Exactitud
Árbol de Decisión	51.02%
KNN	90.06%
SVM	83.05%
Análisis Discriminante	43.70%
Redes Neuronales	71.40%
Promedio	67.84%

Tabla 30. Resultados Primera Fase

5.3.2 Fase II

En esta segunda fase, se proponen los coeficientes de Chebyshev, Fourier, y los Momentos de Comb como características vectoriales de la imagen, para mejorar la exactitud de las técnicas que se utilizaron y poder comparar los resultados. Como se muestra en la tabla 31, se obtuvo un índice de precisión promedio del 80.28 %, mejorando significativamente el valor anterior.

Técnica utilizada	Exactitud
Árbol de Decisión	58.87%
KNN	90.15%
SVM	83.05%
Análisis Discriminante	81.36%
Redes Neuronales	75.12%
Promedio	80.28%

Tabla 31. Resultados Segunda Fase

5.3.3 Fase III

En esta fase, se consideraron las 7 clases principales del Sistema de Harvard [2], así como también los Coeficientes de Chebyshev, Fourier, y los Momentos de Comb. Obteniendo un índice de precisión promedio del 75.47%. Los resultados se muestran en la tabla 32.

Técnica utilizada	Exactitud
Árbol de Decisión	50.25%
KNN	90.32%
SVM	75.78%
Análisis Discriminante	65.51%
Redes Neuronales	89.14%
Promedio	75.47%

Tabla 32. Resultados Tercera Fase

5.3.4 Fase IV

Para terminar, en esta última fase se consideraron las clases B, A, F, G y K con sus respectivas subclases. De igual forma, se utilizaron las técnicas de aprendizaje automático, como son: árboles de decisión, K-NN, la máquina de soporte vectorial (SVM), el análisis discriminante y redes neuronales (RN), así como también los Coeficientes de Chebyshev, Fourier, y los Momentos de Comb. Obteniendo un índice de precisión promedio del 78.22%, como se muestran en la tabla 33.

Técnica utilizada	Exactitud
Árbol de Decisión	64.93%
KNN	90.24%
SVM	75.78%
Análisis Discriminante	73.97%
Redes Neuronales	86.19%
Promedio	78.22%

Tabla 33. Resultados cuarta fase

5.3.5 Resultados de las fases propuestas

Se muestra en la tabla 34, las tasas de exactitud alcanzadas en las cuatros fases propuestas. Se concluye, que se obtuvieron mejores resultados en la segunda fase, en la cual se fusionaron dos clases de acuerdo a su temperatura. Las demás fases, no alcanzaron altos índices de exactitud, debido a que las clases se encuentran desbalanceadas. Se concluye que el conjunto de datos de firmas espectrales, alcanzó una exactitud del 80.28% en la fase II. Los resultados finales indican que la técnica K-NN obtuvo la clasificación más alta en las cuatro fases con un 90.06%, 90.15%, 90.32 %, 90.24% de exactitud. Se logró clasificar con éxito las cinco y siete clases por tipo espectral. Finalmente, con estos resultados se demuestra que la técnica K-NN obtiene el mejor resultado en comparación con las demás técnicas utilizadas.

Fase	Exactitud
I	67.84%
II	80.28%
III	75.47%
IV	78.22%

Tabla 34. Resultados de las cuatro fases propuestas

5.4 Comparación del método propuesto

El método de clasificación estelar que se probó en este trabajo, proporcionará una alternativa para las tareas de clasificación de imágenes astronómicas y/o satelitales con base en la firma espectral de los objetos. En comparación con [4], donde se logran clasificar cinco de las siete clases analizadas, con un 88.1% de precisión, utilizando la técnica de máquinas de soporte vectorial (SVM), en este trabajo de tesis se logró alcanzar un 90.32 % con la técnica del vecino más cercano (K-NN).

De acuerdo con [40], existen catálogos de espectrometría que contienen parámetros específicos de espectros estelares, son una recopilación de datos observados durante 20 años en diferentes partes del mundo, tales como Chile, Armenia, Tadjikistan y Bolivia. En [4] utilizaron el catálogo de Pulkovo, mencionan que los espectros estelares son representados dada una señal Y de entrada, y utilizan un conjunto de señales tanto para el entrenamiento como para las pruebas.

A diferencia de [4], una de las ventajas que se obtuvo en este trabajo de tesis es contar con un conjunto de datos de firmas espectrales extraído del acervo de imágenes astronómicas del INAOE. Se concluye que la clasificación presentada en [24], se basó en las características principales de la imagen (firma espectral) y se clasificó el mayor número posible de espectros estelares.

Capítulo 6

Conclusiones

6.1 Conclusiones

En resumen, se utilizaron operaciones de procesamiento de imagen para extraer la información de las imágenes astronómicas, se presentaron técnicas de aprendizaje automático para realizar la clasificación de espectros estelares, realizando 4 experimentos con parámetros propuestos que se basaron en los coeficientes de Chebyshev, Fourier, Wavelet y los Momentos de Comb. Posteriormente se trabajó con la clasificación del sistema de Harvard, logrando clasificar cinco [B, A, F, G, K] y posteriormente siete clases [O, B, A, F, G, K, M]. Asimismo, se clasificaron 4 subclases de cinco clases establecidas. En relación con lo anterior, se obtuvo como resultado un conjunto de datos de firmas espectrales, y se demostró que al utilizar técnicas de aprendizaje automático y parámetros adecuados, se puede alcanzar un alto nivel de exactitud y precisión en la clasificación. De esta forma, se concluye que para ambos casos de clasificación y sub clasificación, la técnica de aprendizaje automático con la mayor tasa de exactitud fue el vecino más cercano (K-NN) para todos los casos, logrando cumplir con los objetivos planteados en este trabajo de tesis.

6.2 Contribuciones

Las principales contribuciones de este trabajo se enlistan a continuación:

- Se obtuvieron métodos automáticos, que clasificaron un mayor número de clases y subclases, con respecto a trabajos previos.

- Se propuso un conjunto de datos de firmas espectrales extraído del acervo de imágenes digitales, obtenido de algoritmos de extracción y selección de características.

6.3 Publicaciones

- P. Avila Guzmán, A. E. Ortiz Esquivel, R. Díaz Hernández y L. Robles Altamirano, «Stellar Spectra Classification with Machine Learning Classifier Systems and Chebyshev Coefficients using the Tonantzintla Schmidt Camera Plate Collection,» de *IEE Conference Publication Services*, Guadalajara, Jalisco, 2019.
- P. Avila Guzmán, A. E. Ortiz Esquivel, R. Díaz Hernández y L. Robles Altamirano, «Stellar Spectra Classification with Machine Learning Classifier Systems and Chebyshev Coefficients using the Tonantzintla Schmidt Camera Plate Collection,» de *Mexican International Conference on Artificial Intelligence*, Guadalajara, Jalisco, 2019.

6.4 Trabajo futuro

Para trabajo futuro, se propone evaluar otras técnicas de aprendizaje automático y se plantea crear un método para identificar y extraer de manera automática los espectros estelares en toda la imagen astronómica (se considera que una imagen astronómica contiene hasta 300 sub imágenes y requiere un trabajo arduo para la recolección de muestras de espectros estelares) con el fin de obtener un mayor número de muestras para las clases de tipo O y M. De acuerdo con Castilla Carrillo en [7], una de las aplicaciones más importantes de las firmas espectrales es la clasificación, gracias a las características análogas que posee un objeto es más sencillo categorizarlo en clases. Esto indica, gran aplicabilidad en el área de geociencias (ciencias de la tierra), para clasificar el agua, la vegetación y el tipo de suelo. En este sentido, se propone extrapolar el método propuesto en este trabajo de tesis, utilizando imágenes satelitales.

Referencias

- [1] J. L. Castilla Carrillo, «IPICIM: Módulo Clasificador de Imágenes Ópticas Multiespectrales Aplicado al área de Geociencias,» San Luis Potosí, 2016.
- [2] A. Meadows, «Evolucion estelar,» *Reverte*, 1986.
- [3] «INAOE,» [En línea]. Available: <http://inaoep.mx/>. [Último acceso: 20 Mayo 2018].
- [4] H. P.-B. J. d. J. R.-M. J. M. R.-C. a. L. A.-R. Cruz-Vega, «Automatic stellar spectral classification with multiple intelligent classifiers,» *IEEE International Instrumentation and Measurement Technology Conference*, pp. 1-5, 2017.
- [5] R. Díaz Hernández, H. Peregrina Barreto, L. Altamirano Robles, J. González Bernal y A. Ortiz Esquivel, «Automatic stellar spectral classification via sparse representations and dictionary learning,» *Springer Science*, 2014.
- [6] S. Haykin, «A comprehensive foundation, Neural Network,» p. 41, 2004.
- [7] J. L. C. Carrillo, «IPICIM: Módulo Clasificador de Imágenes Ópticas Multiespectrales Aplicado al área de Geociencias,» San Luis Potosí, 2016.
- [8] J. Hernández y D. Montaner, «Patrones de respuesta espectral,» vol. 1, p. 14, 2007.
- [9] H. K. Resnick, Física Vol. 2 5ta ed., México: Continental, S.A de C.V, 1999.
- [10] J. L. Castilla Carrillo, «IPICIM: Módulo Clasificador de Imágenes Ópticas Multiespectrales Aplicado al área de Geociencias,» San Luis Potosí, 2016.
- [11] «Elementos químicos,» [En línea]. Available: <http://herramientas.educa.madrid.org/tabla/espectros/spespectro.html>. [Último acceso: 18 Febrero 2018].
- [12] NASA, «Radioastronomía,» [En línea]. Available: https://www.mdscn.nasa.gov/index.php?Section=Lineas_de_Absorcion_y_de_Emission. [Último acceso: 23 Febrero 2018].
- [13] A. Meadows, «Evolucion estelar,» *Reverte*, 1986.
- [14] J. P. A. Moya, *Procesamiento y Análisis de imágenes digitales*, Costa Rica, 2012.
- [15] «Procesamiento Digital de Imágenes,» [En línea]. Available: <http://dea.unsj.edu.ar/imagenes/recursos/Capitulo1.pdf>. [Último acceso: 12 Mayo 2018].
- [16] M. M. Ortiz, *Procesamiento digital de imágenes*, Puebla, 2013.

- [17] P. J. C. Torres, *Clasificador de señales espectrales astronómicas*, Puebla, 2016.
- [18] «Realce de imagen,» [En línea]. Available: <http://poseidon.tel.uva.es/~carlos/ltif10001/realce.pdf>. [Último acceso: 16 Mayo 2018].
- [19] V. Uhlmann, S. Singh y A. Carpenter, «CP-CHARM: Segmentation-free image classification made accessible,» *BMC Bioinformatics*, p. 19, 2016.
- [20] N. Orlov, L. Shamir, T. Macura, J. Hohnston y I. Golberg, «Wndchrm – an open source utility for biological image analysis,» *Springer*, 2008.
- [21] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. Eckley y I. Goldberg, «WND-CHARM: Multi-purpose image classification using compound image transforms. Pattern Recognition Letters,» *ResearchGate*, vol. 29, nº 11, 2008.
- [22] F. Samaria y A. Harter, «Parameterisation of a stochastic model for human face identification,» de *Workshop on Applications of Computer Vision*, 1994.
- [23] L. Shamir, «Automatic morphological classification of galaxy images,» *Monthly Notices of the Royal Astronomical Society*, 2009.
- [24] A. P. Avila Guzmán, A. E. Ortiz Esquivel, R. Díaz Hernández y L. Robles Altamirano, «Stellar Spectra Classification with Machine Learning Classifier Systems and Chebyshev Coefficients using the Tonantzintla Schmidt Camera Plate Collection,» de *Mexican International Conference on Artificial Intelligence*, Guadalajara, Jalisco, 2019.
- [25] L. S. T. M. J. J. D. M. Nikita Orlov, «WND-CHARM: Multi-purpose image classification,» *Pattern Recognition Letters*, Vols. %1 de %2 ISSN 0167-8655, pp. 1684-1693, 2008.
- [26] J. Boyd, «Chebyshev y de Fourier métodos espectrales,» *Chebyshev y de Fourier métodos espectrales. Courier Corporation*, 2001.
- [27] Gradshtein, I. Tabla Ryzhik de integrales, series y los productos (5ª ed.), Academic Press, 1994.
- [28] I. Gradshtein y I. Ryzhik, «Table of integrals, series and products,» *Academic Press*, vol. 5, p. 1054, 1994.
- [29] J. P. Boyd, «Chebyshev and Fourier spectral methods,» *Courier Corporation*, 2001.
- [30] R. W. Y. S. Z. Ping, «Descripción de la imagen con momentos Chebyshev-Fourier,» *J. Opt. Soc. Am. A*, pp. 1748-1754, 2002.
- [31] K. y. R. L. Kira, «El problema de selección de características: métodos tradicionales y un nuevo algoritmo,» pp. Actas AAAI-92., 1992.
- [32] Matlab, «Machine Learning,» [En línea]. Available:

- https://la.mathworks.com/discovery/machine-learning.html?s_tid=srchtitle.
[Último acceso: 20 Febrero 2018].
- [33] Matlab, «SVM,» [En línea]. Available: <https://la.mathworks.com/discovery/svm-maquina-vectores-soporte.html>. [Último acceso: Marzo 26 2018].
- [34] Matlab, «K-NN,» [En línea]. Available: <https://la.mathworks.com/help/stats/classification-nearest-neighbors.html>. [Último acceso: 3 Abril 2018].
- [35] H. Abdi, « "Análisis de correspondencia discriminante",» *Enciclopedia de medición y estadística . Thousand Oaks (CA):* , pp. 270-275., 2007.
- [36] J. L. C. Carrillo, «Módulo Clasificador de Imágenes Ópticas Multiespectrales Aplicado al área de Geociencias,» San Luis Potosí, 2016.
- [37] J. A. C. Ochoa, «Técnicas de validación,» Instituto Nacional de Astrofísica, Óptica y Electrónica, [En línea]. Available: <https://ccc.inaoep.mx/~ariel/TecnicasDeValidacion2011.pdf>. [Último acceso: 15 Agosto 2018].
- [38] J. F. H, « Cross-Validation Explained,» de *Institute for Genomics and Bioinformatics*, 2005.
- [39] R. E. Corrales Andino, «Catalogo de firmas espectrales de especies florales en la ciudad universitaria Tegucigalpa,» Honduras, 2012.
- [40] A. A. A. V. D. G. E. I. H.-T. I. N. V. N. V. N. V. P. E. R. D. S. G. A Alekseeva, «THE PULKOVO SPECTROPHOTOMETRIC CATALOG OF BRIGHT STARS IN THE RANGE FORM 320 TO 1080 NM,» ST. PETERSBURG, RUSSIA, 1996.
- [41] L. C. J., «MNRAS,» pp. 389, 1179, 2008.
- [42] Díaz Hernández, Raquel, «“Digitalización de placas astronómicas con microdensitometría”,» 2000.
- [43] R. E. W. R. C. González, «Digital Image Processing,» New Jersey, 2002.
- [44] J. C. Torres Patiño, *Clasificador de señales espectrales astronómicas*, Puebla, 2016.
- [45] R. F. Guindos,
<https://w3.ual.es/~fguindos/IMtdi/Tratamiento%20Digital%20de%20Im%E1genes%20con%20IMtdi.pdf>, Almería, 2001.
- [46] A. D. Torres, *Procesamiento digital de imágenes*, Perfiles Educativos, 1996.
- [47] R. W. Y. S. Z. Ping, «Image description with Chebyshev–Fourier moments,» *J. Opt. Soc*, pp. 1748-1754, 2002.
- [48] Á. M. F. d. Cía, «Conservación de Archivos fotográficos,» 15 Marzo 2018. [En línea].

Available: <http://loc.gov/preservation/care/photoleaspanish.html>..

- [49] G. McLachlan, «Análisis discriminatorio y reconocimiento estadístico de patrones,» *Wiley Interscience*, pp. ISBN 0-471-69115-1 . MR 1190469 ., 2004.
- [50] A. H. C. S. G. y. G. J. Bahramian, «Stellar encounter rate in galactic globular clusters.,» *The Astrophysical Journal*, vol. 2, nº 136, p. 766, 2013.
- [51] «Poster- RN,» [En línea]. Available: <https://www.microsiervos.com/archivo/ordenadores/poster-redes-neuronales.html>. [Último acceso: 4 Junio 2018].
- [52] S. A. Martín, *Procesamiento Digital de Imágenes*, España: Editorial Academia Española, 2011.
- [53] Matlab, «Matlab,» [En línea]. Available: <https://la.mathworks.com/products/matlab.html>. [Último acceso: 27 Julio 2018].
- [54] EduSpace, «Teledetección,» [En línea]. Available: http://www.esa.int/SPECIALS/Eduspace_ES/SEM6DYD3GXF_o.html. [Último acceso: 12 Junio 2018].
- [55] M. Li, «A Review of Remote Sensing Image Classification Techniques: The Role of Spatio contextual Information.,» *European Journal of Remote Sensing*, pp. 389-411, 2014.
- [56] H. R. Díaz, «Análisis Espectrofotométrico de las Placas Astronómicas de la Cámara Schmidt de Tonantzintla», 2005.
- [57] R. Díaz Hernández, H. Peregrina Barreto, L. Altamirano Robles, J. González Bernal y A. Ortiz Esquivel , «Automatic stellar spectral classification via sparse representations and dictionary learning,» *Springer Science*, 2014.