Universidad de Deusto

Caracterización de Nanomateriales mediante Tratamiento de Imagen, Reconstrucción 3D y Técnicas de IA

Juan José López de Uralde Huarte

Bilbao, marzo de 2013

Universidad de Deusto

Caracterización de Nanomateriales mediante Tratamiento de Imagen, Reconstrucción 3D y Técnicas de IA

Tesis doctoral presentada por
Juan José López de Uralde Huarte
dentro del Programa de Doctorado en
Sistemas de Información

Dirigida por el Dr. D. Pablo García Bringas
y la Dra. Dña. Teresa Guraya Díez

El doctorando | El director | La directora

Bilbao, marzo de 2013

Resumen

La nanotecnología no es sólo la ciencia del futuro, sino lo es también del presente. Se utiliza en todos los sectores, desde la sanidad a la energía pasando por las tecnologías de la información y el transporte.

Para la presente investigación, hemos tomado como caso de uso el *carbon black*, un nanomaterial mezclado con multitud de materiales para mejorar sus propiedades, como la resistencia a la abrasión o al envejecimiento en neumáticos y plásticos, o la fuerza de los pigmentos entre otras aplicaciones.

En la actualidad, la industria analiza los nanomateriales mayoritariamente mediante métodos indirectos, que observan el cambio en su estado a medida que se les añade aceite o hidrógeno. De esta forma, se estima el área superficial y se calculan algunos indicadores que se relacionan con las propiedades del material.

No obstante, nosotros hemos optado por mejorar los métodos directos existentes, que consisten en analizar los nanomateriales a partir de imágenes de microscopio. Hemos avanzado, tanto en el tratamiento de imagen como en las características extraídas. De hecho, algunas de éstas han superado a las características existentes en la literatura.

Además, hemos empleado por primera vez el aprendizaje automático para la categorización de agregados. De esta forma, se identifica automáticamente su morfología, lo que determinará las propiedades finales del material con el que sea mezclado.

Por último, hemos presentado un algoritmo genético de reconstrucción de agregados a partir de sólo dos imágenes ortogonales, por medio del cual se extrae más información que con una tomografía, la cual requiere de un número elevado de imágenes.

Resumiendo, hemos mejorado el estado de la técnica de los métodos de estudio directo, permitiendo en un futuro cercano desbancar a los métodos indirectos utilizados actualmente.

Abstract

Nanotechnology is not only the science of the future, but it is indeed the science of today. It is used in all sectors, from health to energy, including information technologies and transport.

For the present investigation, we have taken carbon black as a use case. This nanomaterial is mixed with a wide variety of materials to improve their properties, like abrasion resistance, tire and plastic wear or tinting strength in pigments.

Nowadays, indirect methods of analysis, like oil absorption or nitrogen adsorption are the most common techniques of the nanomaterial industry. These procedures measure the change in the physical state while adding oil and nitrogen. In this way, the superficial area is estimated and related with the properties of the material.

Nevertheless, we have chosen to improve the existent direct methods, which consist in analysing microscopy images of nanomaterials. We have made progress in the image processing treatments and in the extracted features. In fact, some of them have overcome the existing features in the literature.

In addition, we have applied, for the first time in the literature, machine learning to aggregate categorization. In this way, we identify automatically their morphology, which will determine the final properties of the material that is mixed with.

Finally, we have presented an aggregate reconstruction genetic algorithm that, with only two orthogonal images, provides more information than a tomography, which needs a lot of images.

To summarize, we have improved the state of the art in direct analysing techniques, allowing in the near future the replacement of the current indirect techniques.

Agradecimientos

La presente tesis doctoral ha sido posible gracias al apoyo de un gran número de personas, unas me han ayudado directamente y otras han estado a mi lado a lo largo de este camino.

En primer lugar, quiero agradecer a mi director de tesis, el Dr. Pablo García Bringas la oportunidad de trabajar en el por aquel entonces S^{3}lab, ahora DeustoTech Computing, y de animarme a realizar esta tesis doctoral, por su ayuda y consejos durante estos años. Además, quiero agradecer a Sendoa Rojas Lertxundi por ponerme en contacto con él, sin el que ahora no me encontraría escribiendo estas líneas. También agradezco a Agustín Zubillaga Rego por introducirme en el mundo *Nano*. Y por supuesto agradezco a la Universidad de Deusto y a DeustoTech la oportunidad que me han brindado de seguir formándome a la par que trabajar en investigación.

En segundo lugar, quiero mencionar al resto del extraordinario equipo que es el S^{3}lab, del que puedo sentirme orgulloso de formar parte. Comienzo por Mikel Salazar González, por su disposición a la hora de hacer magníficas ilustraciones para esta tesis y otras publicaciones científicas, además de por haber trabajado en colaboración, también junto a Aitor Santamaría Ibirika en la realización del algoritmo genético de reconstrucción 3D de agregados de *carbón black. Ánimo Mikel con tu tesis y no reinventes la rueda en tu afán de perfección.*

No puedo olvidarme del brillante Dr. Igor Santos Grueiro, el cual me ha guiado desde el primer artículo científico hasta la finalización de esta tesis. Tengo que reconocer su labor tractora para fomentar la investigación en el laboratorio y sacarlo adelante, junto con los tres doctores Javier Nieves Acedo, Borja Sanz García y Carlos Laorden Gómez. Con Jorge de la Peña Sordo, he tenido el placer de programar un robot para cifrar mensajes en un cubo de

Rubik y de jugar numerosos partidos de paddle al acabar la jornada. Asimismo, con Félix Brezo Fernández he compartido amenas conversaciones.

Por otro lado, esta tesis debe mucho al equipo EMERGE, encabezado por mi directora de tesis, Teresa Guraya Díez, por su dedicación y esfuerzo en producir contribuciones de calidad. A Julen Ibarretxe Uriguen, por su pensamiento crítico, a Ana Okariz Larrea, por sus reconstrucciones tomográficas, y a Maider Iturrondobeitia Ellacuria, por sus horas de microscopio, *ánimo con las arcillas*.

Tampoco habría sido posible esta tesis sin la beca recibida del Ministerio de Educación, ni la financiación de la Diputación Foral de Bizkaia, a través de los proyectos: Optimización de técnicas de caracterización tridimensional de nanopartículas y materiales nanoreforzados, CONAN, COMALTO y PROMEC.

Quiero agradecer al Dr. Dierk Hartmann la oportunidad de realizar una estancia de 3 meses en la Universidad de Kempten. A Christian Heidrich, por su gran acogida desde el primer día, así como a Korbinian Seifert y Jens Zimmermann, compañero de despacho y de cervezas. *Vielen Dank für alles.*

Me gustaría mencionar al grupo de amigos de Bilbao, en especial al ya mencionado Javi, amigo inseparable desde la carrera, compañero de trabajo y ahora también compañero de piso. A Igor, por tener siempre las puertas de su casa de Londres abiertas aunque nos presentemos de sorpresa. A los magníficos amigos, Iñigo e Iker, un abrazo y ánimo en estos momentos difíciles. A Sonia, por su alegría contagiosa. A Juan el argentino, por estar siempre ahí. A los fiteranos, los más chulos después de los bilbaínos. A los amigos de conciertos, por los buenos momentos pasados y futuros. Y a Bea y a Pol, por lo feliz que me habéis hecho.

Por último, termino agradeciendo a mi familia, tanto a mis padres como a mi hermana, por todo el apoyo recibido y el gran ejemplo que han supuesto para mí. Por no acabar con mi afición de abrir cualquier aparato para ver cómo funcionaba y terminar sacándole los motores.

Por último, un agradecimiento a los ausentes. Si hubiera nombrado a todos los que se merecen, los aquí presentes habrían perdido protagonismo, así que espero no haber ofendido a nadie. *Gracias a todos.*

Índice general

Índice general **vii**

Índice de figuras **xiii**

Índice de tablas **xv**

1 Introducción **1**

1.1 Los nanomateriales 2

1.1.1 Proceso de mezclado 3

1.1.2 Propiedades y aplicaciones 4

1.1.3 Nanomaterial elegido 5

1.2 Retos a superar 6

1.3 Hipótesis 8

1.4 Objetivos 9

1.5 Solución propuesta 12

1.6 Metodología de investigación 14

1.7 Estructura del documento 15

2 Caracterización del *carbon black* **17**

2.1 El *carbon black* 18

2.1.1 Fabricación del *carbon black* 19

2.1.2 Clasificación del proceso de producción del *carbon black* 21

2.1.2.1 Proceso de oxidación térmica 22

2.1.2.2 Proceso de descomposición térmica 23
2.1.3 Materias primas 23
2.1.4 Post-tratamiento del *carbon black* 24
2.1.4.1 Post-tratamiento por oxidación 24
2.1.4.2 Post-tratamiento con vapor de agua 24
2.1.5 Caracterización 25
2.1.5.1 Métodos indirectos 25
2.1.5.2 Métodos directos 26
2.1.6 Morfología 28
2.2 Tratamiento de imagen 29
2.2.1 Tipos de microscopios 30
2.2.1.1 SEM 31
2.2.1.2 TEM 33
2.2.1.3 AFM 34
2.2.2 Estándar ASTM 35
2.2.3 Eliminación de ruido 36
2.2.4 *Binarización* 38
2.2.4.1 Umbral óptimo de la escala de grises 38
2.2.4.2 Detección de bordes 42
2.2.5 Segmentación 47
2.2.5.1 Operaciones morfológicas 47
2.2.5.2 Esqueletización 48
2.3 Extracción de características 49
2.3.1 Fase I: *Binarización* 51
2.3.2 Fase II: Filtrado de ruido 52
2.3.3 Fase III: Identificación de agregados 54
2.3.4 Fase IV: Extracción de características 55
2.4 Conjuntos de datos 63
2.4.1 SEM 64
2.4.2 TEM 66
2.4.3 Artificiales 67

2.5 Selección de atributos 67
2.5.1 Clasificación de los métodos 68
2.5.2 Métodos utilizados 70
2.5.3 Conclusiones 72
2.6 Evaluación empírica 73
2.6.1 Conjunto SEM 74
2.6.2 Conjunto TEM 78
2.6.3 Conjunto artificial 82
2.6.4 Tiempos 86
2.7 Discusión de los resultados 86
2.8 Sumario 88

3 Categorización del *carbon black* 91
3.1 Introducción 92
3.2 Aprendizaje Automático 93
3.2.1 Redes Bayesianas 94
3.2.2 *Support Vector Machines* 95
3.2.3 *K-nearest-neighbours* 97
3.2.4 Árboles de decisión 98
3.3 Análisis ROC 100
3.4 Evaluación experimental 105
3.4.1 Metodología general 105
3.4.2 Conjunto SEM 108
3.4.3 Conjunto TEM 111
3.4.4 Conjunto artificial 114
3.5 Discusión de los resultados 117
3.6 Sumario 121

4 Reconstrucción 3D 123
4.1 Introducción 124
4.2 Reconstrucción superficial 125
4.3 Reconstrucciones tomográficas 126

4.4 Algoritmos de Montecarlo 127
4.5 Algoritmo RandomSalt 128
4.5.1 Segmentación del agregado 129
4.5.2 Mapa de alturas 130
4.5.2.1 Preprocesado de imagen 130
4.5.2.2 Eliminación del ruido ambiente 131
4.5.2.3 Obtención del valor de una partícula 131
4.5.2.4 Generación del mapa de alturas 132
4.5.3 Algoritmo genético 132
4.5.3.1 Inicialización 133
4.5.3.2 Mutaciones 133
4.5.3.3 Selección 134
4.5.3.4 Finalización 135
4.5.4 Reconstrucción tridimensional y visualización 137
4.6 Evaluación empírica 138
4.7 Discusión de los resultados 142
4.8 Sumario 143

5 Conclusiones **145**
5.1 Síntesis de la validación del sistema 146
5.2 Resumen de los resultados obtenidos 149
5.3 Aplicaciones de la investigación 152
5.4 Limitaciones de la solución 153
5.5 Trabajo futuro 154
5.6 Consideraciones finales 156

6 Conclusions **157**
6.1 Synthesis 157
6.2 Summary of results 160
6.3 Research applications 163
6.4 Limitations of the model 164
6.5 Future Work 165

6.6 Final considerations 167

Publicaciones **169**

Bibliografía **171**

Índice de figuras

1.1 Esquema de la arquitectura. 13

2.1 Tamaño de partícula, agregado y aglomerado de *carbon black*. 19
2.2 Categorías morfológicas de los agregados de *carbon black*. . . . 27
2.3 Influencia del tamaño de partícula y de la estructura del *carbon black* en sus propiedades. 29
2.4 Esquema funcionamiento microscopios SEM y TEM. 30
2.5 *Carbon black* capturado con un microscopio SEM. 31
2.6 Microscopio SEM Hitachi - S4800. 32
2.7 *Carbon black* capturado con un microscopio TEM. 32
2.8 Microscopio TEM Philips - EM208S. 33
2.9 *Carbon black* capturado con un microscopio AFM. 34
2.10 Microscopio AFM Digital Instruments - Nanoscope IIIa. . . . 35
2.11 Máscaras de convolución. 44
2.12 Aplicación de operadores primera derivada. 46
2.13 Representación del operador LG 47
2.14 *Convex hull* de un agregado de *carbon black*. 48
2.15 Esqueletización de un agregado de *carbon black*. 49
2.16 Algoritmo de extracción de características. 50
2.17 Elemento estructural con forma de disco para aplicar operaciones de dilatado y erosión 53
2.18 Imagen SEM original y *binarizada* 55

2.19 Área media de los agregados por cada tipo de morfología para el conjunto SEM. 56
2.20 Sonda ultrasónica Bioblock Scientific Vibracell 75043. 64
2.21 Imagen SEM de muestra. 65
2.22 Imagen TEM de muestra. 66
2.23 Imagen artificial de muestra. 67

3.1 Ejemplo de una red bayesiana. 94
3.2 Ejemplo de un SVM bidimensional. 96
3.3 Ejemplo de un clasificador KNN. 98
3.4 Ejemplo de un árbol de decisión. 100
3.5 Funcionamiento curva ROC. 104

4.1 Reconstrucción superficial a partir de un mapa de relieve. . . . 125
4.2 Agregado de *carbon black* capturado mediante microscopio TEM a -45º y a +45 º. 128
4.3 Segmentación del agregado de *carbon black* capturado a -45º y a +45º para ser utilizado como máscara. 129
4.4 Mapa de alturas del agregado de *carbon black* capturado a -45º y a +45º. 131
4.5 Proyecciones de las partículas a los planos correspondientes de las imágenes originales. 134
4.6 Representación basada en esferas y representación volumétrica fusionada con la representación de *isosuperficie*. 137
4.7 Evolución del *fitness* con respecto a las iteraciones. 140
4.8 Agregado capturado a 0º, su mapa de alturas y una proyección. 141
4.9 Modelo 3D de un agregado de *carbon black* 144

Índice de tablas

2.1 Clasificación de los procesos de producción. 21

2.2 Relevancia de los atributos según los métodos Chi-cuadrado y Relación de Ganancia de Información sobre el conjunto de datos SEM. 75

2.3 Relevancia de los atributos según los métodos *ReliefF* y *Random Forest* sobre el conjunto de datos SEM. 77

2.4 Relevancia de los atributos según los métodos Chi-cuadrado y Relación de Ganancia de Información sobre el conjunto de datos TEM. 79

2.5 Relevancia de los atributos según los métodos *ReliefF* y *Random Forest* sobre el conjunto de datos TEM. 81

2.6 Relevancia de los atributos según los métodos Chi-cuadrado y Relación de Ganancia de Información sobre el conjunto de datos artificial. 83

2.7 Relevancia de los atributos según los métodos *ReliefF* y *Random Forest* sobre el conjunto de datos artificial. 85

2.8 Tiempo requerido en segundos en evaluar los atributos con cada uno de los tres conjuntos de datos. 86

3.1 Matriz de confusión. 101

3.2 Resultados de los algoritmos de aprendizaje automático con el conjunto de datos SEM. 108

3.3 Matriz de confusión para el método *Random Forest* de 1.000 árboles aplicado sobre el conjunto de datos SEM. 109

3.4 Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos SEM. 110
3.5 Resultados de los algoritmos de aprendizaje automático con el conjunto de datos TEM. 112
3.6 Matriz de confusión para el método *Random Forest* de 1.000 árboles aplicado sobre el conjunto de datos TEM. 112
3.7 Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos TEM. 113
3.8 Resultados de los algoritmos de aprendizaje automático con el conjunto de datos artificial. 115
3.9 Matriz de confusión para el método *Support Vector Machine* con *kernel* universal *Pearson VII* aplicado sobre el conjunto de datos Artificial. 116
3.10 Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos artificial. 117

«A veces estamos demasiado dispuestos a creer que el presente es el único estado posible de las cosas.»

Marcel Proust (1871-1922)

1

Introducción

A pesar de que pueda parecer que los nanomateriales hacen referencia a un material fabricado mediante técnicas desarrolladas en los últimos años, éstos ya se encuentran en la naturaleza. Es más, se han hallado nanopartículas en los pulmones de un cuerpo humano de más de 5.000 años de antigüedad en los Alpes [HMPP00] y nanotubos de carbono que datan de hace 10.000 años en muestras de hielo extraídas de un glaciar en Groenlandia [EM04].

Más recientemente, se han usado como pigmentos y en la actualidad se emplean en áreas muy diversas, como la ingeniería [GG08], biomedicina [SM09]y bioingeniería [GK08]. En muchos casos se buscan materiales cuya característica principal y razón de uso sea su resistencia mecánica a la fatiga o al desgaste, su capacidad de amortiguar vibraciones, su resistencia a los ataques con productos químicos, sus propiedades eléctricas cuando se refuerzan con partículas conductoras o su facilidad para ser conformados a geometrías complejas. Éstas partículas, poseen una superficie muy elevada en relación a su volumen, lo que hace que esta interacción sea tan particular [Mar04]. Otra característica de los nanomateriales es su capacidad para modificar propiedades fundamentales, como la magnetización, propiedades ópticas o temperatura de fusión respecto a los materiales a escala micro o macroscópicas.

A pesar de las cuantiosas inversiones que se han realizado hasta la fecha para el estudio de los nanomateriales [Ini12, CR11], todavía no se conocen a la

perfección sus mecanismos de formación y las leyes físicas que rigen el modo en el que interactúan con el material con el que son mezclados [MEE05]. En esta investigación, se pretende proporcionar tanto a la industria, como a los laboratorios de investigación de una herramienta para el control de calidad y para el estudio exhaustivo de los nanomateriales.

El resto del capítulo queda organizado de la siguiente forma. La sección 1.1 muestra una visión general del campo de los nanomateriales. El apartado 1.1.1 expone la importancia del proceso de mezclado, el apartado 1.1.2 la forma en que modifican las propiedades del material con el que son mezclados y el apartado 1.1.3 el nanomaterial elegido, el *carbon black* o negro de humo. Posteriormente se plantean los retos a superar en la sección 1.2, es decir, las dificultades existentes para su estudio y los enfoques actuales. A continuación, en la sección 1.3, se expone la hipótesis que ha guiado la realización de la tesis, así como los objetivos que han surgido de ella en la sección 1.4. Adicionalmente, se presenta la solución propuesta en la sección 1.5 y la metodología de investigación seguida en la sección 1.6. Por último, se describe la estructura del documento en la sección 1.7.

1.1 Los nanomateriales

La finalidad de los nanomateriales es la de mejorar las propiedades del material con el que son mezclados. Éste último se denomina matriz, y cada uno de los elementos que forman un compuesto, incluyendo las nanopartículas, la matriz y los aditivos reciben el nombre de fase. Los materiales de relleno se comenzaron a utilizar para expandir el material y abaratar su coste [Mar04]. Además, modifican las propiedades de la matriz. Las partículas de relleno tienen formas y tamaños muy variados, por ejemplo, para los polímeros[1] son necesarios tamaños inferiores a 40 μm. Pero si son menores que 3 μm proporcionan unas mejoras más notables. Las nanopartículas son las que proveen unas propiedades óptimas cuando están bien dispersas.

Específicamente, una nanopartícula tiene una dimensión menor de 100 nm en al menos una dimensión [BM11a]. Sin embargo, nunca se encuentran en solitario y mediante uniones químicas forman lo que se denomina como nanoagregados. De igual modo, los agregados forman aglomerados mediante

[1]Los polímeros son macromoléculas formadas por la unión de moléculas de menor tamaño llamadas monómeros. Estas macromoléculas son generalmente orgánicas, es decir, contienen carbono. El término polímero, se utiliza a veces para referirse a los plásticos, pero en realidad incluye también otras sustancias, como el almidón, la celulosa, la seda o el ADN.

fuerzas de *Van der Waals*[1] [Don98] para conservar su energía interna [Mar04]. Aplicando una suave fuerza mecánica o mediante disolventes, se pueden separar de nuevo en agregados [BRH$^+$06], lo que se suele hacer en el momento de mezclarse con otro material al que se le quieren cambiar sus propiedades.

Se conoce, que la forma de las partículas está directamente relacionada con la manera en que interactúan con la matriz [Mar04]. No obstante, los mecanismos de refuerzo continúan siendo en parte un misterio y no son comprendidos completamente [MEE05].

En la actualidad existen numerosas técnicas para estudiar la morfología de los nanoagregados [Rus07, MG99, MGP05, MGLLP$^+$10], una de las características más usadas es la relación entre la altura y la anchura; o entre la longitud más larga y su anchura. En concreto, una esfera tiene una relación de 1. A medida que la forma de un agregado va de una esfera a un bloque, a una plancha o a un tubo, el valor de esta relación disminuye. Asimismo, el área superficial del nanomaterial aumenta, y es precisamente la extensión en la que interactúa con la matriz.

1.1.1 Proceso de mezclado

El proceso de preparación de la mezcla del nanomaterial con la matriz es de vital importancia. Para obtener un material con unas propiedades mecánicas o eléctricas uniformes es necesario tener en cuenta los 3 elementos presentes: incorporación, distribución y dispersión. Estos 3 fundamentos se dan simultáneamente durante el proceso de mezclado.

- La incorporación implica el mezclado de los diferentes ingredientes en una masa consistente pero todavía heterogénea.
- La homogenización se da en el proceso de distribución en el que el nanomaterial se reparte aleatoriamente dentro de la matriz.
- La dispersión hace referencia a la separación de los aglomerados del material de relleno y a su asociación física. A su vez, está formado por tres etapas: (i) mojado inicial, (ii) ruptura de aglomerados y (iii) empapado completo de las partículas para desplazar las bolsas de aire y mejorar la

[1] Las fuerzas de *Van der Waals* son fuerzas de atracción o repulsión que tienen lugar entre átomos, moléculas o superficies. Son relativamente débiles en comparación con los enlaces químicos normales y son causadas por correlaciones en las polarizaciones fluctuantes de partículas cercanas, una consecuencia de la dinámica cuántica [AGD75].

fusión con la matriz consiguiendo un contacto total entre el nanomaterial y la matriz.

Cuanto menor es el tamaño de las partículas, el proceso de dispersión es más costoso. Asimismo, a menor proporción de relleno, el proceso es más laborioso [Mar04]. En estos casos, a menudo se realiza una mezcla muy concentrada que será fraccionada con una extrusora y posteriormente mezclada de nuevo con la matriz en unas proporciones menores.

1.1.2 Propiedades y aplicaciones

La mejora en las propiedades que un material de relleno, nanomaterial o no, puede proporcionar a la matriz va desde la conductividad a la permeabilidad, pasando por las propiedades mecánicas y ópticas o la resistencia a la abrasión. Ya que las matrices poliméricas son las más usadas, especialmente para el *carbon black*, que es el nanomaterial en el que se centra la tesis, a continuación se explicará la forma en la que influyen los nanomateriales en los polímeros.

A la hora de determinar el compuesto necesario para una aplicación hay que definir las condiciones a las que se verá sometido el material. Una cuestión de gran importancia también es el envejecimiento causado por fatiga o por las condiciones ambientales, por lo que en diferentes partes del mundo es posible que para una misma aplicación se requiera una mezcla diferente [Mar04]. En este aspecto, para matrices de caucho, el *carbon black* consigue alargar la vida del compuesto a diferencia de la sílice, a la cual hay que agregarle antioxidantes para evitar su envejecimiento [Whi01].

La mejora de las propiedades mecánicas es el uso tradicional de los nanomateriales y hoy en día continúa siendo su principal aplicación. Por ejemplo, en el caso de los polímeros, la rigidez, aumenta con la adición de relleno. En el caso de la tracción y elongación hasta el punto de ruptura no siguen siempre el mismo patrón. Para algunos elastómeros[1] blandos, la tracción sí que aumenta con la adición de relleno hasta un punto óptimo [Mor87].

En cuanto a la conductividad térmica de los polímeros, es relativamente baja en comparación con los metales y otros muchos materiales inorgánicos [Mar04]. Esta propiedad se puede incrementar mezclándolos con metales y

[1]Los elastómeros son polímeros elásticos que se encuentran sobre su temperatura de transición vítrea, lo que les proporciona dicha capacidad de deformación. Se usan para cierres herméticos, adhesivos o neumáticos entre otras muchas aplicaciones.

otros rellenos carbonáceos, para ser usados en baterías recargables, dispositivos electrónicos y sensores [AA90]. El punto de fusión de un polímero cristalino apenas se ve afectado por la presencia de rellenos, mientras que la temperatura de transición vítrea se ve ligeramente incrementada. En cambio, el umbral de deformación por calor se ve incrementado normalmente entre 10 y 20 °C [Mar04].

Respecto a las propiedades ópticas, los rellenos como calcita y talco pueden dar color a la matriz polimérica, en cambio, para rellenos con tamaños de partícula inferior a la longitud de onda de la luz, 0.4 μm, la mezcla se vuelve transparente [Mar04].

De la misma forma, los materiales nanocristalinos, nanomateriales cuyas nanopartículas tienen una estructura principalmente cristalina, modifican a su vez las propiedades de la matriz, permitiendo controlar además de las propiedades ópticas su conductividad eléctrica, de gran utilidad en paneles solares [KL04].

Una mayor descripción de las propiedades que se pueden obtener está fuera del alcance de esta tesis por lo que se proponen las siguientes fuentes de información para el lector que esté interesado [Mar04, MEE05, SDE$^+$03, WN09].

1.1.3 Nanomaterial elegido

El *carbon black* es un material que se ha utilizado desde el siglo III a. C. por los chinos, indios y egipcios como pigmento. En el siglo XV, con la invención de la imprenta, la necesidad de un pigmento fuerte incrementó sustancialmente su demanda. En 1912 la empresa Diamond Rubber, descubrió el efecto que produce el *carbon black* al añadirlo al caucho, reforzándolo considerablemente [CH99]. De esta forma, tuvo lugar un avance muy importante para la industria del caucho y, en consecuencia, la del automóvil. En la actualidad, no sólo se utiliza en estos ámbitos, sino que tiene otros muchos usos entre los que se encuentran el revestimiento de materiales [SWKN11] y la creación de materiales absorbentes de radiación [KO11].

Así, el *carbon black* es usado frecuentemente como relleno de elastómeros, plásticos y pinturas para modificar sus propiedades mecánicas, eléctricas y ópticas; para posteriormente establecer sus aplicaciones en un segmento de mercado determinado. Al mezclarlo con plásticos proporciona protección ultravioleta, conductividad eléctrica, rango de oscuridad, opacidad y refuerzo.

Por otro lado, al usarse con caucho aumenta su resistencia a la fractura y a la abrasión. La producción actual supera los 8 millones de toneladas, dedicándose el 90 % al caucho, principalmente para neumáticos, el 9 % como pigmento y el 1 % restante como ingrediente indispensable en una gran variedad de aplicaciones [Ass06c].

Las principales características que influyen en los compuestos de *carbon black* son su (i) tamaño de partícula, (ii) tamaño de agregado, (iii) la morfología de los agregados y su (iv) microestructura [DBW93]. Además también son de vital importancia la naturaleza de su superficie, su organización estructural y porosa, el área de su superficie y su composición química.

Es por tanto clara la necesidad de un conocimiento más profundo de la naturaleza del *carbon black*, como de sus mecanismos de formación, para mejorar los procesos de producción. Por esto, en la exposición del estado de la técnica sobre el *carbon black* en la sección 2.1 se describe su industria, los mecanismos de formación y el modo en el que influyen las diferentes características sobre las propiedades finales del material. Como trabajo futuro a esta tesis, se aplicarán las técnicas creadas, que sobre el *carbon black* se hayan probado exitosas, sobre otro tipo de materiales.

1.2 Retos a superar

Para poder plantear una hipótesis en condiciones de rigurosidad es necesario identificar qué retos existen actualmente y así plantear una línea de trabajo que cubra las necesidades actuales. El área de estudio de los materiales es tan amplia que se ha tomado como caso de estudio el *carbon black*. Un nanomaterial cuyos procesos de producción, a pesar de su madurez, todavía dan cabida a cuantiosa investigación.

La morfología de los nanomateriales, como se ha explicado en la sección 1.1 influye en gran medida en las propiedades finales del material, y es precisamente la dificultad para estudiarla de forma directa en lo que se ha centrado esta tesis.

- **Optimización de la imagen y eliminación de ruido.** En 1931, con la invención de los microscopios electrónicos se superó la limitación de los microscopios ópticos que no permitían en su momento la visión a aumentos superiores de 1.000X. Desde entonces, se ha dedicado y se le

sigue dedicando un gran esfuerzo al estudio de los materiales por medio del análisis de imagen. Estas son obtenidas de microscopios electrónicos, tanto de transmisión como de barrido, así como de microscopios de fuerza atómica. El principal problema es la baja resolución de las imágenes y la cantidad de ruido existente en ellas. Para eliminar el ruido se han utilizado tradicionalmente diferentes filtros como el gaussiano, que tienen el inconveniente de desenfocar la imagen. Sin embargo, existe otro método, la difusión anisotrópica [PM90, BSMH98], con la que definiendo el nivel de suavizado adecuado no se distorsionan los bordes. Esta técnica se ha utilizado con éxito en imágenes a escala nanométrica [SFF03]. Además, en la práctica, no es posible obtener las imágenes siempre bajo las mismas condiciones, por lo que es necesario un método robusto que proporcione unos resultados uniformes y aceptables.

- **Detección de agregados.** Para localizar los agregados existen diferentes técnicas. Las dos ramas más importantes se basan en la detección de bordes [SF68, Can86, Pre70, Rob65] y en la determinación de un umbral [Nob79, ZH08, SS04] de la escala de grises para discernir entre el fondo y los agregados. Ambas pueden realizarse globalmente, utilizando información de toda la imagen; o localmente, utilizando información de sólo una zona de la imagen. La primera de ellas tiene el problema de que, debido a la naturaleza de las imágenes capturadas con microscopios electrónicos, el borde de los agregados a menudo no queda cerrado, lo cual es un requisito indispensable para poder delimitarlo. La otra alternativa, la del umbral, aunque en algunos casos no delimita los bordes con tanta precisión, es mucho más robusta. En este contexto, surge la problemática a abordar, que al igual que en el reto anterior, es la necesidad de métodos adecuados para imágenes no uniformes[RS06]. Así, aunque se intentan capturar en las mismas condiciones, cada microscopio es diferente y existen diferentes variables que influyen en el resultado final, como el estado del filamento de la pistola de electrones.

- **Caracterización y categorización.** Como hemos recalcado, la forma es de vital importancia. Generalmente, la manera de analizar su distribución se diferencia entre estudios directos e indirectos. En la actualidad, se encuentran más extendidos los métodos indirectos, como son la absorción de aceite y la adsorción de nitrógeno [WN09], en parte por su comodidad y rapidez. Estos métodos, absorción y adsorción[1], estudian

[1]La adsorción es la atracción entre la superficie exterior de una partícula sólida con otro

cómo cambia el estado del polvo de *carbon black* a medida que se le añade aceite o nitrógeno, y lo relacionan con sus propiedades físicas. Sin embargo, para esta tesis se han elegido los métodos directos por la gran cantidad de información que se puede extraer con ellos y también para poder conocer la dispersión del nanomaterial en la matriz. Así, además del área superficial, se pueden extraer numerosas características estructurales, como la altura y la anchura [HMH92], la información fractal [Kay84], los diámetros de Feret [Ame07] o información del esqueleto [MG99]. Para clasificar los agregados a partir de las características extraídas, en la literatura se han definido diferentes categorías morfológicas [HMH92]. Para mejorar esta categorización vamos a aumentar el número de características existentes en la literatura y vamos a estudiar nuevos métodos de clasificación.

- **Reconstrucción 3D.** Para tener una información más completa y precisa de los nanomateriales se realizan normalmente reconstrucciones tomográficas a partir de numerosas proyecciones [HGM+00]. Otro enfoque es el de la reconstrucción geométrica a partir de vistas múltiples basada en las restricciones de la geometría *epipolar*, tradicionalmente usada con pares de imágenes para extraer información tridimensional de una parte del objeto [RS06]. Estos métodos requieren mucho trabajo manual para la captura y ajuste de las imágenes y es necesario automatizarlos completamente o desarrollar un algoritmo que no precise de tantas imágenes para poder extraer información y conclusiones a partir de un número suficiente de muestras.

1.3 Hipótesis

En esta sección se va a definir la hipótesis que guiará la realización de la tesis en un afán de ponerla a prueba. Para llegar a esta hipótesis se han identificado primero los retos existentes en el área de los nanomateriales y específicamente en el caso del *carbon black* en la sección 1.2. Una vez definida la hipótesis en base al problema de partida existente, en la sección 1.4 se define el objetivo principal y los objetivos específicos y operacionales que surgen de él para la validación de la hipótesis.

elemento, mientras que la absorción es la retención del elemento en la estructura física del sólido.

Así, la hipótesis de partida es:

> *Es posible desarrollar un sistema automatizado para el control de la calidad de nanomateriales formados por negro de humo y para la mejora de su proceso productivo, por medio del tratamiento de imágenes de microscopio a través de su caracterización y posterior aplicación de técnicas de machine learning.*

El cumplimiento de la hipótesis expuesta da solución a un procedimiento realizado mayoritariamente con medidas indirectas, o mediante el análisis de imágenes de forma manual o semi-automática, lo que es costoso y no permite en muchos casos extraer unas conclusiones suficientemente significativas debido al tamaño de la muestra que es posible analizar en un tiempo razonable. Además, también tiene como objetivo el poder ser de asistencia en los procesos de producción de nanomateriales con nuevas propiedades. Para la aprobación de esta hipótesis habrá que conseguir que el sistema sea completamente autónomo y que permita una categorización con unas tasas de acierto y AUC (Area Under the ROC Curve)[1] competentes. Asimismo, los nuevos parámetros de caracterización que se definan tienen que suponer un incremento sustancial en la cantidad de información extraída de los agregados.

1.4 Objetivos

Una vez expuesta la hipótesis se identifica a continuación el objetivo principal que surge con la intención de probarla:

> **Objetivo principal 1** *Desarrollar y evaluar un caracterizador y categorizador de agregados de negro de humo que permita realizar un control de calidad de nanomateriales y asista en los procesos de producción de materiales con nuevas propiedades.*

[1]El área por debajo de la curva ROC es un índice relativo a la probabilidad de que un clasificador acierte que una instancia pertenece a una clase en vez de que suceda un falso positivo, es decir, que sea clasificado incorrectamente como de dicha clase. Para clasificaciones no binarias[2], como es nuestro caso, es la media balanceada del AUC de cada clase.

[2]Una clasificación no binaria es aquella que tiene más de dos clases o categorías.

De este objetivo principal se desglosan una serie de objetivos específicos:

Objetivo específico 1 *Desarrollar y evaluar un caracterizador de agregados de negro de humo por medio de técnicas de selección de atributos.*

Objetivo específico 2 *Desarrollar y evaluar un categorizador de agregados de negro de humo por medio de técnicas de machine learning.*

Objetivo específico 3 *Desarrollar y evaluar un reconstructor 3D de agregados que permita estudiar su estructura real y obtener características adicionales.*

El primer objetivo incluye tanto el desarrollo de nuevas características como su validación. Esta validación se llevará a cabo de dos maneras: primero mediante el análisis de ellas con diferentes algoritmos de selección de atributos y posteriormente, una vez que se cree el categorizador morfológico, por medio del estudio de la mejora de la clasificación que suponga la adición de estas nuevas características a las ya establecidas en la literatura. En el tercer objetivo, se plantea la reconstrucción tridimensional de un agregado a partir de sólo dos imágenes TEM. Este método será una alternativa a las laboriosas tomografías que se realizan en la actualidad para estudiar la estructura real de un nanoagregado. Esta parte se validará comparándolas con las tomografías en términos de volumen, área superficial y afinidad con una proyección intermedia a las dos utilizadas para la reconstrucción. Además, con la caracterización de la reconstrucción se podrá alimentar en un futuro al categorizador de agregados en los casos que se priorice la precisión frente a la velocidad y coste de personal capturando imágenes.

Para llevar a cabo los objetivos específicos recién planteados es necesario marcar unos objetivos operacionales que permitan lograrlos:

Objetivo operacional 1 *Diseñar e implementar un sistema que realice un tratamiento sobre una imagen de microscopio electrónico que permita extraer información relevante de ella.*

Objetivo operacional 2 *Diseñar e implementar un sistema que extraiga información de una imagen que contenga un agregado segmentado binarizado*[1]*.*

Objetivo operacional 3 *Diseñar e implementar un sistema que asista en el etiquetado manual de agregados.*

Objetivo operacional 4 *Diseñar e implementar un sistema que clasifique agregados por medio de algoritmos de machine learning usando las características extraídas de los agregados.*

Objetivo operacional 5 *Validar el categorizador desarrollado incluyendo un estudio de la relevancia de las características morfológicas.*

Objetivo operacional 6 *Diseñar e implementar un sistema que sea capaz de reconstruir tridimensionalmente un agregado por medio de sólo dos imágenes TEM, así como validarlo con reconstrucciones tomográficas.*

[1]La segmentación es el proceso de delimitación de los agregados y la binarización es el proceso de transformación de una imagen en escala de grises a una imagen con solo dos valores posibles para cada píxel, negro o blanco, correspondientes al fondo y al agregado respectivamente.

Objetivo operacional 7 *Diseñar e implementar un sistema que obtenga nuevas características de la reconstrucción tridimensional.*

Objetivo operacional 8 *Diseñar e implementar un sistema que permita realizar controles de calidad de manera automatizada por medio de la categorización de agregados.*

Objetivo operacional 9 *Diseñar e implementar un sistema que permita analizar las características de un material para el control de la producción.*

Con la consecución de los objetivos expuestos la hipótesis planteada quedará demostrada, obteniendo un sistema que servirá tanto para realizar controles de calidad de nanomateriales como para asistir en los procesos de producción de nuevos materiales.

1.5 Solución propuesta

Tras exponer los objetivos que dirigen la presente investigación, a continuación, se presenta la arquitectura de la solución desarrollada. Ésta contiene los distintos componentes que se describen en el capítulo 2, relativo a la extracción de información de las imágenes, en el capítulo 3 se realiza una categorización morfológica en base a la información extraída y en el capítulo 4 se expone el método de reconstrucción 3D desarrollado.

En la Figura 1.1 se muestra la solución creada, la cual se compone de los siguientes módulos:

- **Detección de agregados.** En este módulo las imágenes capturadas mediante un microscopio electrónico son procesadas. Para comenzar, se elimina el ruido de la imagen. Posteriormente, se aplican varios filtros y operaciones a la imagen para detectar los agregados. Este proceso se conoce como segmentación.

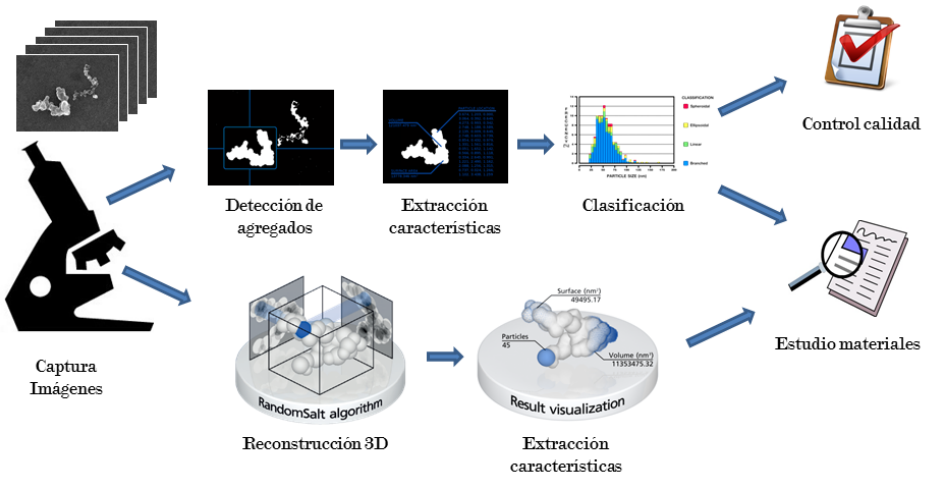


Figura 1.1: Esquema de la arquitectura.

- **Extracción de características.** Un vez segmentados los agregados, se extraen de ellos el área y perímetro, con los que además se estiman otros valores, además se realizan diversas mediciones, como los diámetros de Feret[1] o se trazan segmentos aleatoriamente para analizar sus intersecciones con el perímetro del agregado.

- **Clasificación de la morfología.** También conocido como categorización, es el módulo en el que, a partir de las características extraídas, se asigna una de las cuatro categorías definidas por Herd [HMH92] de forma automática mediante algoritmos de aprendizaje automático.

- **Reconstrucción 3D.** Este módulo lleva a cabo una reconstrucción tridimensional de un agregado a partir de únicamente dos imágenes tomadas con 90º de diferencia. Mediante un algoritmo genético, sobre varias soluciones posibles, de forma iterativa, se realizan mutaciones aleatorias con las que se van aproximando a la estructura real del agregado. De esta forma, se consigue obtener un agregado mucho más similar al real que caracterizando agregados bidimensionales, sin tener que realizar una costosa reconstrucción tomográfica.

- **Extracción características del modelo 3D.** Al igual que en el segundo módulo, se extraen características, pero esta vez de un modelo

[1]Un diámetro de Feret es el valor de la distancia entre dos paralelas perpendiculares a una dirección fija, y tangentes a la silueta proyectada del agregado. En nuestro caso calculamos el máximo, el mínimo y el perpendicular al máximo.

tridimensional. Esta información extraída consiste en el área superficial, el volumen, el número de partículas y el tamaño de éstas.

A partir de la información extraída en los módulos anteriores, se realizarán dos informes con el propósito de llevar a cabo controles de calidad o estudiar las características de diferentes materiales.

- **Informe de categorías.** Con la finalidad de asistir en el control de calidad a personal no experto, se provee de un informe claro, con la distribución de cada categoría para el material analizado.
- **Comparativa de características.** En el caso de querer comparar diferentes materiales por personal experto, se proporciona un informe completo de las distribuciones de las categorías, así como de todas las características extraídas para diferentes materiales. De esta forma se podrán observar con facilidad las diferencias existentes en los materiales al variar los procesos de producción. Asimismo, se podrán estudiar los cambios morfológicos que sufre un nanomaterial tras el proceso de mezclado comparando las características de los agregados antes y después de ser mezclados.

1.6 Metodología de investigación

Para la realización de la presente investigación se ha seguido la metodología de trabajo que a continuación se expone. Así, se definen los pasos a seguir en el orden presentado, no obstante, esto no implica que, una vez superado un punto, se dé por finalizado, ya que con seguridad será necesario volver a adquirir nuevo conocimiento y desarrollar nuevas soluciones que serán validadas y difundidas repetidamente.

1. **Adquisición del conocimiento:** En esta etapa, no sólo se estudiará el estado de la técnica en el área de interés directa, sino que se identificarán áreas relacionadas cuyo conocimiento pueda aportar nuevas soluciones a los problemas existentes en el ámbito sobre el que se plantea la hipótesis. Por ello se aplicarán algoritmos de *machine learning* a la categorización de agregados y se estudiará el tratamiento de imagen tanto a nivel general como las técnicas utilizadas concretamente para las imágenes de microscopios electrónicos.

2. **Diseño y desarrollo de aplicaciones:** A partir de las diferentes técnicas identificadas se elegirán las más adecuadas para diseñar y desarrollar una aplicación que las implemente. No se elegirá una solución única y cerrada sino que consistirá en probar diferentes enfoques y combinaciones de éstos.

3. **Evaluación:** En esta fase se evaluará el prototipo desarrollado por medio de experimentos para las diferentes técnicas utilizadas, como serán los diferentes algoritmos de *machine learning*, diferentes técnicas de tratamiento de imagen y sobre diferentes tipos de *carbon black* con diferentes tipos de microscopios electrónicos.

4. **Análisis:** Una vez concluida la experimentación se procederá a analizar los resultados obtenidos extrayendo conclusiones a partir de éstos y de los conocimientos adquiridos previamente.

5. **Difusión de resultados:** Las conclusiones obtenidas de los experimentos serán publicadas tanto para ser evaluadas por la comunidad científica como para obtener una respuesta de esta que ayude a mejorar la investigación. Esta respuesta llegará de diversas formas: *feedback* de los revisores de los artículos publicados, cuestiones planteadas en las presentaciones de los artículos y referencias a los artículos publicados. Este contacto con la comunidad científica puede crear incluso vínculos que contribuyan a la investigación de manera importante, como por ejemplo, por medio de la realización de proyectos en colaboración o estancias en centros de investigación especializados en el área de los materiales, tratamiento de imagen y/o inteligencia artificial en este caso.

1.7 Estructura del documento

A partir de la introducción, el documento contiene tres capítulos de contribuciones, cada uno de ellos expone inicialmente el estado de la técnica que forma la base de experimentos llevados a cabo y termina con las conclusiones obtenidas sobre el trabajo realizado. En concreto, el contenido de cada capítulo es el siguiente:

- **Capítulo 1. Introducción.** En este capítulo, se presenta el potencial de los nanomateriales y, en concreto, el de los polímeros reforzados con negro de humo. Se exponen sus propiedades y se plantean los retos existentes en esta área. Posteriormente, se enuncia la hipótesis y se dan a

conocer los objetivos que surgen de ella. A continuación, se explica la arquitectura de la solución propuesta. Asimismo, se explica la metodología seguida que garantiza el rigor de la investigación.

- **Capítulo 2. Caracterización del *carbon black*.** En este capítulo se da a conocer el proceso de extracción de características de agregados de *carbon black*. Se comienza exponiendo el estado de la técnica de este material, incluyendo los diferentes métodos de fabricación y caracterización. Adicionalmente, se describen las técnicas existentes para el tratamiento de imagen necesarias para su procesado. Posteriormente, se explican los pasos seguidos para la extracción de características y se introducen atributos novedosos. Además, se explica el estado de la técnica en el ámbito de la selección de atributos y se describen en profundidad los cuatro métodos elegidos para la evaluación de los atributos. Así, es posible apreciar la relevancia que tienen para discernir entre las cuatro clases morfológicas.

- **Capítulo 3. Categorización del *carbon black*.** En este capítulo, se expone el estado de la técnica empleado para la clasificación automática mediante métodos de *machine learning* así como del análisis ROC de los resultados. Además, se evalúa el método de categorización en cuatro categorías morfológicas a partir de las características extraidas en el capítulo anterior con tres conjuntos de datos. Asimismo, se exponen los esfuerzos realizados mediante muestras sintéticas para mejorar los resultados balanceando las diferentes clases y ampliando el número de instancias.

- **Capítulo 4. Reconstrucción 3D.** En este capítulo, se describen las diferentes alternativas para reconstruir objetos tridimensionalmente. Se exponen las tomografías y un método basado en las simulaciones de Montecarlo y se comparan con el algoritmo propuesto. Éste, es un algoritmo genético que sólo precisa dos imágenes ortogonales. Adicionalmente se presenta un método de validación con una tercera imagen.

- **Capítulo 5. Conclusiones y trabajo futuro.** En este capítulo, se exponen las conclusiones, presentando las publicaciones realizadas. A continuación, se plantean las posibles aplicaciones de los algoritmos desarrollados en esta tesis. Para finalizar, se expone el trabajo futuro identificado.

«La materia siempre existe en conjunción con la forma que la caracteriza, no puede existir por sí sola, es un elemento de todos los cuerpos.»

Aristóteles (384-322 a. C.)

2

Caracterización del *carbon black*

Los nanomateriales y en concreto el *carbon black* se llevan estudiando desde hace muchos años, pero es en la última década en la que se ha dado un mayor crecimiento en las inversiones realizadas para su investigación y mejora. En este capítulo, se exponen tanto las técnicas más tradicionales, como las más recientes e innovadoras para el análisis morfológico del *carbon black*. Asimismo, se presenta el estado de la técnica que forma la base para el desarrollo de la presente investigación, incluyendo diferentes métodos para el tratamiento de imagen y las técnicas más relevantes de selección de atributos.

A pesar de los grandes avances de las últimas décadas en la capacidad de magnificación de los microscopios, hay un problema inherente a la técnica, que es el ruido existente en sus imágenes [VC07]. Los métodos de tratamiento de imagen tienen el problema de no ser robustos frente a las variables condiciones de captura. Así, aunque se intentan replicar, los pequeños cambios influyen en los parámetros optimizados para el tratamiento de unas imágenes en concreto.

Por otra parte, de estas imágenes optimizadas se extrae información muy variable, como el perímetro fractal [Kay84], los diámetros de Feret [Ame07] o características del esqueleto [MG99]. En concreto, las contribuciones que hemos realizado en este ámbito son las siguientes:

- Creación de un método robusto para la segmentación de agregados de *carbon black* en imágenes capturadas con diferentes microscopios y condiciones.
- Definición de nuevas características para extraer de los agregados.
- Validación de las nuevas características y comparación con las existentes en la literatura.

El resto del capítulo queda organizado como sigue. La sección 2.1 describe los diferentes métodos de producción del *carbon black* y las materias primas que requiere. La sección 2.2 expone las diferentes técnicas existentes para el procesamiento de imágenes y la extracción de información de ellas, centrándose en las imágenes de microscopio pero describiendo también los algoritmos empleados en otros tipos de imágenes. La sección 2.3 explica las cuatro fases de las que se compone el algoritmo de caracterización. A continuación, en la sección 2.4 se describen los conjuntos de datos utilizados en el presente capítulo para la extracción de características y en el capítulo 3, sobre la categorización del *carbon black*. Posteriormente, en la sección 2.5 se exponen los diferentes métodos de selección de atributos existentes, explicando en mayor profundidad los métodos utilizados. Seguidamente, en la sección 2.6 se evalúan tanto los nuevos atributos como los existentes en la literatura. A continuación, en la sección 2.7 se discuten los resultados expuestos, y por último, en la sección 2.8 se resumen las contribuciones de este capítulo.

2.1 El *carbon black*

El negro de humo o *carbon black* es un material producido por la combustión incompleta de los productos derivados del petróleo. Es una forma de carbono amorfo con una relación superficie-volumen extremadamente alta y que, como tal, es uno de los primeros nanomateriales en utilizarse a gran escala. Sus dos usos principales son, como pigmento y como refuerzo en productos de goma y plástico [Ass06a].

Sus partículas elementales son esferoidales, aunque éstas nunca se encuentran en solitario, sino que forman agregados [Don98], por lo que éstos pueden ser considerados como la verdadera unidad principal [HMSH93]. Siguiendo las fuerzas de *Van der Waals*, los agregados se conectan formando aglomerados. La Figura 2.1 muestra una representación gráfica del tamaño de las partículas, agregados y aglomerados.

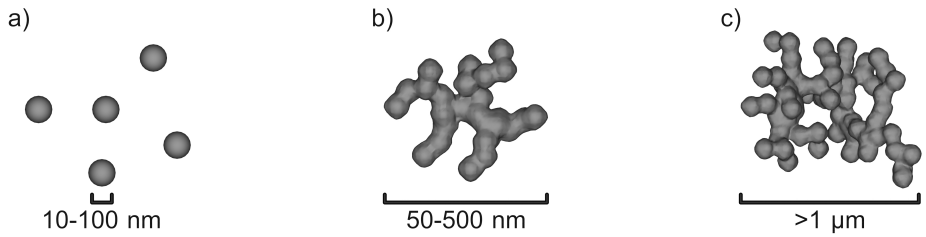


Figura 2.1: *Carbon black*: a) partículas, b) agregado y c) aglomerado.

Normalmente se produce en forma de polvo y se mezcla con otros materiales, proporcionándoles una mayor resistencia a la tracción, rotura y abrasión. Estos cambios son atribuidos a las características moleculares, químicas y reológicas[1] del elastómero, de las características del relleno y del proceso y tecnología de mezclado [FNL05].

La producción mundial actual de *carbon black* supera las 8 toneladas [Ass06c], sin embargo, todavía no se conocen con total precisión sus mecanismos de formación [MEE05, Leb10]. A continuación, en la sección 2.1.1, se explica brevemente el proceso de producción del negro de humo, y en la sección 2.1.2 se detallan los dos métodos básicos de producción. Más tarde, en la sección 2.1.3, se mencionan las materias primas utilizadas para la producción del *carbon black*. En la sección 2.1.4 se explican los tratamientos posteriores que se aplican al *carbon black* una vez fabricado, para mejorar sus cualidades. En la sección 2.1.5 se describen los dos tipos de caracterizaciones que se pueden realizar a este material. Por último, en la sección 2.1.6, se explica la relevancia de la morfología de los agregados y cómo influye en las propiedades finales del material mejorado con *carbon black*.

2.1.1 Fabricación del *carbon black*

Durante miles de años, el uso del *carbon black* era sólo posible una vez descubierto un proceso de producción adecuado. Es bien sabido que la restricción de oxígeno a las llamas de aceites o resinas ardiendo produce materiales carbonosos [DBW93]. Y es precisamente este hecho la base de los procesos de producción históricos.

En la actualidad la mayoría de la producción del *carbon black* se realiza basándose en el proceso de combustión incompleta de hidrocarburos. Sin em-

[1]La reología es el estudio de la forma en la que fluye la materia. Mediante *reómetros* se somete al material a deformaciones para medir el esfuerzo necesario para realizarlas.

bargo, existe otro proceso llamado descomposición térmica, durante el cual el *carbon black* se forma en ausencia de oxígeno. Estos dos métodos sirven para clasificar inicialmente los métodos de producción existentes, en el apartado 2.1.2, se detallan y subdividen en nuevas clasificaciones.

Cabe resaltar que en el presente estudio son de interés exclusivamente los procesos de fabricación controlados de manera precisa por técnicas de medición y control que permiten la producción de *carbon black* con propiedades claramente definidas. Esto es debido a que se pretende incidir en estos procesos, estableciendo además unos controles de calidad automáticos y eficientes. Por lo tanto, no se investigarán los procesos en los que se libera hollín como un subproducto contaminado, como en calderas mal calibradas o durante la quema descontrolada de materiales carbonosos, como la madera, carbón o aceite [DBW93].

En principio, una planta industrial para la producción a gran escala de *carbon black* dispone de las siguientes secciones:

1. Instalaciones para el almacenamiento de la materia prima
2. Unidades de producción de *carbon black*
3. Equipamiento para la separación del *carbon black* de los gases liberados
4. Procesamiento final del *carbon black*
5. Instalaciones para el almacenamiento del producto final
6. Aprovechamiento de los gases de desecho

Cada sección está interconectada por sistemas de transporte en entornos completamente cerrados para evitar la liberación de *carbon black* en los alrededores. Esto, normalmente no es por cuestiones de salud, ya que hay estudios que indican que la exposición prolongada al *carbon black* no se ha podido relacionar con el riesgo de padecer cáncer de pulmón [SHvT$^+$01, WWN$^+$06]. Aun así, la asociación IARC (International Agency for Research on Cancer), la clasifica como sustancia posiblemente cancerígena para los humanos, pero se les achaca que las condiciones en las que dicha organización realizó los estudios [IAR96] no eran adecuadas y que los resultados obtenidos con ratas eran específicos de dicha especie [Ass06b]. En cualquier caso, en los últimos años se están incrementando las precauciones tomadas en las plantas de producción, en parte también por su poder colorante, por lo que es considerado como un polvo molesto.

2.1.2 Clasificación del proceso de producción del *carbon black*

Desde el punto de vista de la química, es esencial diferenciar entre las dos formas de producir el *carbon black*, que como se han mencionado anteriormente, son la combustión incompleta y la descomposición térmica de hidrocarburos, dependiendo de la presencia o ausencia de oxígeno [VCvSW86].

Tabla 2.1: Clasificación de los procesos de producción.

Proceso químico	**Proceso de producción**	**Materia prima**
Descomposición por oxidación térmica		
Sistema cerrado (flujo turbulento)	Proceso de horno de negro	Aceites aromáticos basados en alquitrán de hulla, petróleo crudo, gas natural
	Proceso de hollín negro	Aceites aromáticos basados en alquitrán de hulla, petróleo crudo
Sistema abierto (llamas de difusión)	Proceso de gas negro Degussa	Destilados de alquitrán de hulla
	Proceso de canal negro (histórico)	Gas natural
Descomposición térmica		
Discontinua	Proceso térmico de negro	Gas natural (aceites)
Continua	Proceso de acetileno negro	Acetileno

El proceso de la combustión incompleta, denominado descomposición por oxidación térmica es con mucho el más importante [WCJ04] y con el que se produce prácticamente la totalidad del *carbon black* mundialmente [DBW93, Leb10]. En términos cuantitativos, el proceso de la descomposición térmica en ausencia de oxígeno es poco significativo. Los métodos usados para la producción del *carbon black* se indican en la Tabla 2.1.

2.1.2.1 Proceso de oxidación térmica

El proceso de oxidación térmica puede ser subdividido dependiendo de si el *carbon black* se produce mediante un flujo turbulento[1] o mediante llamas de difusión[2].

Desde el punto de vista de la fabricación actual, el más relevante es el proceso de flujo turbulento de horno de negro, el cual usa aceites aromáticos como materia prima principal. El horno tiene un reactor cerrado para pulverizar el aceite en condiciones controladas de presión y temperatura. Éste es introducido en una corriente de gas caliente, obtenida mediante una materia prima secundaria que puede ser gas natural o aceite, donde se vaporiza y posteriormente piroliza para formar partículas microscópicas de carbono. Normalmente, la velocidad de reacción se controla con vapor o pulverizadores de agua. Por último, el *carbon black* producido es enfriado y filtrado. Los gases residuales generados son utilizados para generar calor, vapor o energía eléctrica [Ass06c].

Para comprender mejor la diferencia entre los sistemas abiertos y cerrados, podemos fijarnos en las llamas de elementos cotidianos. Un ejemplo de una llama turbulenta son las calderas de aceite de la calefacción central de las casas, mientras que un ejemplo de llama de difusión es una vela. En este último caso se puede apreciar como la llama se divide en diferentes capas en función del oxígeno que dispone. En la capa exterior, donde dispone de suficiente oxígeno, el material carbonoso arde casi por completo. Al mismo tiempo, el calor generado funde y vaporiza la cera. La siguiente capa de la llama, en cambio, dispone de insuficiente oxígeno. Al ser menor la tasa de oxígeno disponible a la reacción de descomposición, se forma carbón, lo que hace resplandecer a la llama. Este carbón al alcanzar la zona exterior en la que hay suficiente oxígeno arderá. No obstante, en caso de situar un objeto frío, como puede ser un vaso, en la llama el carbón formado en la capa interior quedará depositado sobre su base dejándola ennegrecida.

Este proceso es un sistema abierto ya que el oxígeno de los alrededores tiene acceso a la llama y la falta de oxígeno es sólo local y temporal. La formación del *carbon black* en un entorno de corrientes turbulentas requiere un sistema cerrado con un apropiado reactor de flujo. Esto se debe a que de esta forma los componentes necesarios para la reacción pueden introducirse de manera separada y controlada, proporcionándole al sistema una gran flexibili-

[1] Se denomina flujo turbulento al movimiento de un fluido que se da de forma caótica.

[2] Las llamas de difusión se producen cuando el combustible y el oxígeno no se encuentran pre-mezclados y son generalmente amarillas debido a la incandescencia del carbón.

dad. Además, un sistema cerrado evita la pérdida de *carbon black* al exterior, evitando la contaminación.

El proceso denominado sistema abierto de gas negro, proviene del Instituto Alemán de Separación de Oro y Plata dedicado al sector químico y metalúrgico, Degussa (Deutsche Gold und Silber Scheide Anstalt). Por otro lado, el proceso de canal negro es un método histórico de producción de *carbon black* ya en desuso [DBW93].

2.1.2.2 Proceso de descomposición térmica

El proceso de descomposición térmica es similar al de oxidación térmica. El más relevante es el proceso discontinuo que utiliza normalmente gas natural como materia prima. Se denomina proceso térmico de negro y consiste en dos hornos que alternan cada 5 minutos entre precalentamiento y producción de *carbon black*. El gas natural se inyecta en el horno a alta temperatura, y en ausencia de aire se transforma en *carbon black* e hidrógeno [Ass06c].

Se diferencia del proceso de horno de negro principalmente en las siguientes características [WCJ04]:

- El proceso es cíclico en vez de continuo.
- El *carbon black* se forma en ausencia de oxígeno.
- Los gases generados como consecuencia del proceso son prácticamente hidrógeno puro, el cual requiere un procesamiento más complejo que los generados por el proceso de oxidación térmica, que son principalmente nitrógeno y agua.
- El *carbon black* tiene un área superficial mucho menor y una estructura más simple.

Como consecuencia de este último punto, el *carbon black* generado por medio de esta técnica, no tiene tanta capacidad reforzante, ni colorante.

2.1.3 Materias primas

Como ya se ha mencionado anteriormente, los hidrocarburos son la materia prima necesaria para la producción del *carbon black*. Por la forma de las partículas de *carbon black* puede decirse que se forman en la fase gaseosa. Las

partículas de *carbon black* se parecen a otros productos obtenidos por medio de procesos pirogénicos, como es el caso del humo de sílice. Por lo tanto, un prerrequisito de una materia prima para crear *carbon black* es que sea posible transformarla completamente al estado gaseoso [Leb10]. Así, la materia prima utilizada son gases o líquidos que puedan ser vaporizados bajo las condiciones necesarias.

2.1.4 Post-tratamiento del *carbon black*

Existen propiedades deseadas en el *carbon black* que no se consiguen directamente en el proceso inicial de fabricación. Por esto, muchas veces hay que aplicar un tratamiento posterior. Por ejemplo, el *carbon black* usado para dar color a revestimientos de alta calidad debe poseer superficies altamente polares para conseguir una humectación óptima con la sustancia ligante [DBW93]. Los tóners de impresora requieren *carbon blacks* con una gran fuerza colorante, que puedan impartir tanto una conductividad baja como alta a la resina a niveles de concentración relativamente altos. Asimismo, los *carbon blacks* conductores deben tener superficies libres de óxidos y de materiales orgánicos. Estos métodos se basan en la oxidación 2.1.4.1 o en el vapor de agua 2.1.4.2.

2.1.4.1 Post-tratamiento por oxidación

Los *carbon blacks* se comportan de manera diferente ante la oxidación dependiendo de su origen. Además, el material resultante variará en función del proceso de fabricación. Por una parte, el *carbon black* producido con horno sólo contiene pequeñas cantidades de oxígeno en la forma de oxidaciones básicas en la superficie. Por otra, el *carbon black* producido por el método de gas negro Degussa, siempre se encuentra ligeramente oxidado y contiene predominantemente oxidaciones ácidas de la superficie.

2.1.4.2 Post-tratamiento con vapor de agua

El tratamiento posterior del *carbon black* con otra intención diferente a la oxidación se realiza con vapor de agua. Si el post-tratamiento se realiza a temperaturas relativamente bajas (300-500ºC), elimina la materia extraíble de la superficie del *carbon black*. A temperaturas más altas (900-1100ºC) no sólo se agrupa el oxígeno y la materia orgánica se elimina, sino que el *carbon black* también es atacado, lo que le hace poroso.

- **Eliminación de la materia extraíble.** Algunos usos del *carbon black* requieren que los niveles de materia extraíble sean extremadamente bajos, ya que ésta proporciona a veces un color demasiado fuerte. Además, en plásticos blancos y negros o blancos y colorados, pueden aparecer manchas. En estos casos, y en los de materiales que van a entrar en contacto con la comida, el nivel de materia extraíble tiene que cumplir unos límites muy exigentes [Eur84].

- **Producción de *carbon black* poroso.** El *carbon black* poroso puede obtenerse directamente mediante el proceso de producción en horno permaneciendo el tiempo preciso a la temperatura necesaria. No obstante, es aconsejable separar este proceso y hacerlo con vapor, ya que así se crea un proceso independiente del de producción manteniendo las condiciones idóneas para hacerlo poroso. Para obtener *carbon black* poroso en tiempos razonables, la temperatura ronda los 900 - 1.000ºC.

2.1.5 Caracterización

Para caracterizar el *carbon black*, y los nanomateriales en general, existen dos enfoques: los métodos directos y los indirectos. Estos últimos, se llevan a cabo sólo sobre el polvo del nanomaterial, en cambio, los directos, también pueden utilizarse para caracterizar el nanomaterial una vez que está mezclado con la matriz. En el presente trabajo doctoral, uno de los objetivos consiste en desarrollar una metodología que permita estudiar los cambios morfológicos que sufre un nanomaterial tras el proceso de mezclado. Por esto, en el apartado 2.1.5.1 tan sólo se da una pequeña descripción de los métodos indirectos y en la sección 2.1.5.2 se describe de forma más extensa los métodos directos, que son los que se van a emplear.

2.1.5.1 Métodos indirectos

Los dos métodos indirectos más comunes para caracterizar el *carbon black* son la absorción de aceite y la adsorción de nitrógeno [WN09]. La absorción es la retención de un elemento, en este caso el aceite, en la estructura física del sólido, en este caso los agregados de *carbon black*. El número de absorción de aceite de un nanomaterial está directamente relacionado con las propiedades que tendrán los compuestos en los que forme parte [Ame11]. El *absorciómetro*, mide el par de torsión de una muestra de polvo del nanomaterial a la que se le va añadiendo aceite a un ritmo continuo. A medida que la muestra absorbe

el aceite, pasa por tres fases, en la primera fluye libremente, en la segunda se aglomera y en la fase final alcanza la saturación, pasando primero de líquido a sólido y posteriormente de solido a liquido [But03].

La adsorción, en cambio, es la atracción entre la superficie exterior de una partícula sólida, en este caso, el *carbon black*, con otro elemento, en este caso, el nitrógeno. El estándar más utilizado para medir esta propiedad es el método BET [BET38]. La adsorción de N_2 se calcula mediante un equipo capaz de medir volumétrica o gravimétricamente la cantidad de moléculas de este gas que son adsorbidas como una capa por el material estudiado. De esta forma, se obtiene el área superficial externa, que se define como la parte del área superficial que es accesible a la matriz [Ame10].

2.1.5.2 Métodos directos

Los métodos directos se basan en el análisis morfológico de imágenes de microscopio. Comúnmente, los agregados se dividen en cuatro tipos morfológicos diferentes [HMH92], como se muestra en la Figura 2.2. Para discernir entre las cuatro categorías, se calcula la relación *largura/anchura* y la irregularidad del agregado, de todas formas, este método no es trivial ya que incluye un valor difícil de medir: la irregularidad [HMH92] y además, no marca un categoría para los agregados con una relación *largura/anchura* ente 1,5 y 2. Las formas quedan clasificadas en las siguientes categorías:

- **Esferoidal:** Agregados con una relación *largura/anchura* menor que 1,5 pueden ser clasificados como esferoidales.
- **Elipsoidal:** Agregados con una relación *largura/anchura* menor que entre 2 y 3,5 pueden ser clasificados como elipsoidales.
- **Lineal:** Los agregados lineales tienen un ratio *largura/anchura* mayor de 3,5 y baja irregularidad, debido a tener cadenas alargadas con pocas ramas.
- **Ramificada:** Los agregados ramificados tienen también un ratio *largura/anchura* mayor de 3,5 pero con irregularidad elevada, debido a tener más ramas.

Como más tarde se expondrá en la fase de extracción de características, en el apartado 2.3.4, en la literatura existen numerosos parámetros para describir la irregularidad, entre los que se encuentra la dimensión fractal, y la esqueletización, que se describen a continuación.

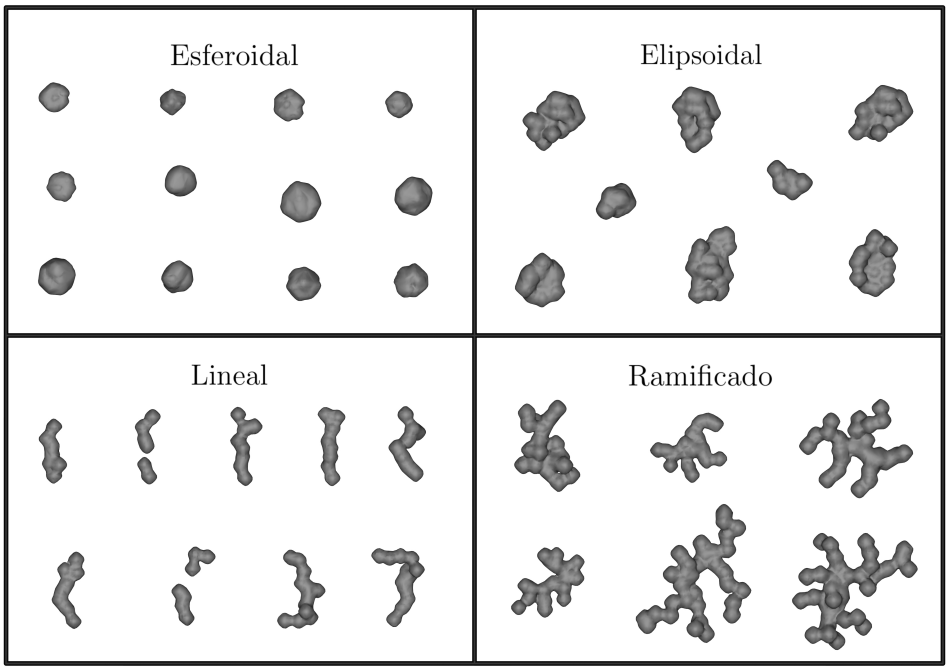


Figura 2.2: Categorías morfológicas de los agregados de *carbon black*.

Dimensión fractal

La dimensión fractal, describe la estructura de los agregados [Mea83]. Concretamente, un fractal es una forma que puede ser dividida en copias más pequeñas de sí mismo [Man82]. Con este fin, Kaye fue el primero en aplicar el análisis fractal a los agregados de *carbon black* [Kay84]. Determinó un *perímetro fractal* basado en el método de Mandelbrot [Man77], de tal forma que, un círculo proporcionaba un valor de $D_p = 1$ y este valor aumentaba a medida que un agregado es más irregular, hasta un máximo de 2. Por otra parte, Gerspacher y O'Farrell [GO91], utilizaron la relación perímetro-área mostrada en la ecuación (2.1) para analizar 15 tipos de *carbon black*:

$$P \sim A^{D_p/2} \tag{2.1}$$

donde P se define como el perímetro del agregado proyectado, A como el área de la proyección y D_p como el perímetro fractal. A mayor irregularidad, mayor es el D_p, sin embargo, las partículas altamente aciculares, es decir muy poco circulares, con un perímetro suave, pueden dar también un alto fractal [DBW93]. Por lo tanto, este método es considerado como una tosca aproximación y se pueden obtener mejores resultados con el algoritmo de

contar cajas [Vos86]. Una *dimensión de caja* de un conjunto S contenido en N es d para cualquier $\epsilon > 0$ siendo $N_\epsilon(S)$ el número menor de cubos n-dimensionales con lados de longitud ϵ necesaria para cubrir S. Así, establece que existe un valor de d que satisface la ecuación (2.2):

$$N_\epsilon(S) \sim 1/\epsilon^d \quad as \quad \epsilon \to 0 \tag{2.2}$$

donde d es la dimensión de S si y sólo si existe una constante positiva k que hace que se cumpla la ecuación (2.3):

$$\lim_{\epsilon \to 0} \frac{N_\epsilon(S)}{1/\epsilon^d} = k \tag{2.3}$$

Esqueletización

Con el fin de clasificar los agregados en los 4 tipos descritos, se realiza un proceso de esqueletización, que será descrito en la sección sobre el tratamiento de imagen en el apartado 2.2.5.2. Por medio de esta técnica, se extraen el número de ramas de un agregado y su longitud [HMSH93, MG99].

2.1.6 Morfología

Como ya se ha explicado, las principales características que influyen en los compuestos de *carbon black* son su (i) tamaño de partícula, (ii) el tamaño de los agregados, (iii) la morfología de los agregados y su (iv) microestructura [DBW93]. Así, en la Figura 2.3, puede apreciarse en qué modo influyen el tamaño de partícula y la estructura de los agregados en la facilidad para su dispersión y en las propiedades finales del material mejorado con *carbon black*.

La facilidad para la dispersión, explicada brevemente en el apartado 1.1.1 sobre el proceso de mezclado de los nanomateriales, es una característica de vital importancia, ya que influirá en el coste de dicho proceso. Dependiendo de las propiedades que se quieran obtener, se requerirá una mezcla más o menos homogénea. En concreto, si se le quiere proporcionar conductividad a un material, una mezcla no uniforme no conseguirá el objetivo propuesto [SSA$^+$91]. Así, la dispersión es más sencilla a mayor tamaño de partícula así como cuanto más compleja es su estructura. En cambio, la conductividad, es mayor a menor tamaño de partícula y cuanto más compleja es su estructura. En cuanto a la capacidad de absorción de aceite, viscosidad y conductividad eléctrica aumentan a menor tamaño de partícula y a mayor complejidad estructural.

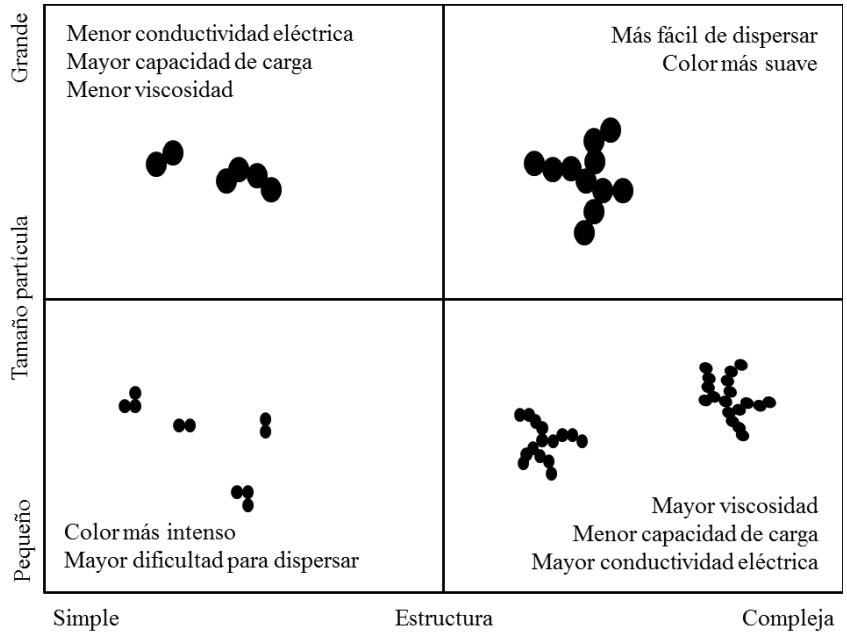


Figura 2.3: Influencia del tamaño de partícula y de la estructura del *carbon black* en sus propiedades.

Por otra parte, la tonalidad de primer plano (del inglés, *masstone*), es más oscura a menor tamaño de partícula, pero no se ve afectada por la estructura. Lo mismo ocurre con el poder colorante que aumenta sólo a menor tamaño de partícula, y con la tonalidad de fondo (del inglés *tinting undertone*), que tiende hacia el marrón a menor tamaño de partícula y hacia el azul en caso contrario. De la misma forma hay otras propiedades que sólo se ven afectadas por la estructura de los agregados. Éstas son el brillo y la capacidad de carga, que aumentan a menor estructura, y el color, que se vuelve ligeramente más fuerte y marrón a menor estructura, y ligeramente más débil y azul a mayor complejidad de su estructura.

2.2 Tratamiento de imagen

Resulta imposible separar completamente la caracterización del *carbon black* del tratamiento de imagen, así, en el apartado 2.1.5 se da una pequeña descripción de las técnicas concretas utilizadas para extraer información del *carbon black*. En cambio, en la presente sección se da una visión más general de los diferentes pasos que se dan desde que se obtiene una imagen hasta extraer información de ella.

Lo que resta de sección queda organizado de la siguiente forma. En el apartado 2.2.1, se comienza exponiendo los diferentes tipos de microscopios empleados para los experimentos. A continuación, en el apartado 2.2.2, se describe brevemente el estándar ASTM para el análisis de imágenes de *carbon black*. Posteriormente, en el apartado 2.2.3, se define el ruido desde el punto de vista de las imágenes y se exponen diferentes métodos para reducirlo. Seguidamente, en el apartado 2.2.4 se presentan los diferentes métodos de *binarización* existentes en la literatura. Por último, en el apartado 2.2.5 se explica el proceso de delimitación de regiones, conocido como segmentación.

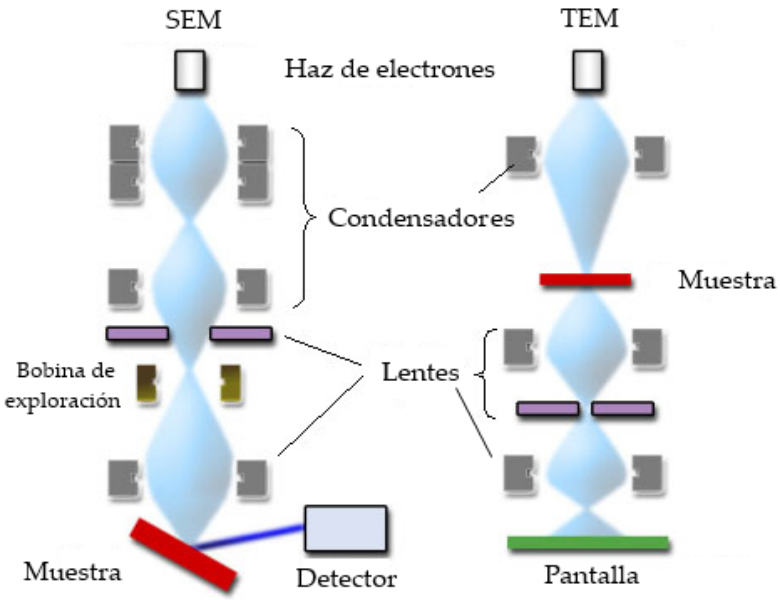


Figura 2.4: Esquema funcionamiento microscopios SEM y TEM.

2.2.1 Tipos de microscopios

Teniendo en cuenta la escala de los agregados de *carbon black*, se necesitan microscopios electrónicos para analizarlos. Destacan dos tipos de microscopios electrónicos: los de transmisión TEM y los de barrido SEM. En la Figura 2.4 se muestra el esquema de funcionamiento de ambos. TEM es una técnica en la que un haz de electrones interactúa con el espécimen al pasar a través de él. Las ondas eléctricas que salen de la muestra son usadas para formar una imagen. Por otro lado, SEM es una técnica en la que una pistola de electrones emite un haz con alta carga eléctrica. Estos electrones viajan a través de

varias lentes magnéticas que enfocan a los electrones hacia un punto muy pequeño. Enfocando este punto a lo largo del espécimen, se construye una imagen a partir de los electrones secundarios que salen despedidos cuando otros electrones chocan contra la superficie de la muestra.

A continuación, se explican los fundamentos de estos dos tipos de microscopios así como del AFM, microscopio probado inicialmente pero descartado por la dificultad para extraer información de imágenes obtenidas con este tipo de microscopio.

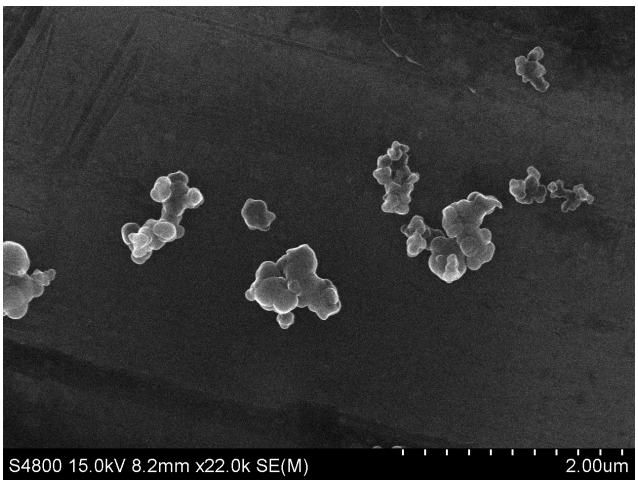


Figura 2.5: Agregados de *carbon black* capturados con un microscopio SEM.

2.2.1.1 **SEM**

Los microscopios SEM (Microscopio Electrónico de Barrido) utilizan un haz de electrones que al incidir en la muestra desprenden electrones secundarios. Los electrones principales y secundarios son capturados para formar la imagen [Ech09]. Los electrones se aceleran previamente en un campo eléctrico para aprovechar su comportamiento ondulatorio. Las muestras sensibles, como las biológicas, se analizan con un voltaje pequeño, en cambio, las muestras metálicas se analizan con un voltaje alto. En este último caso, se aprovecha la menor longitud de onda para obtener una mayor resolución.

En la Figura 2.5 se muestran varios agregados de *carbon black* capturados con un microscopio SEM. Las muestras son normalmente recubiertas con carbón u oro para hacerlas conductoras [Ech09], sin embargo, el *carbon black* ya es conductor, por lo que este paso no es necesario en nuestro material de estudio.

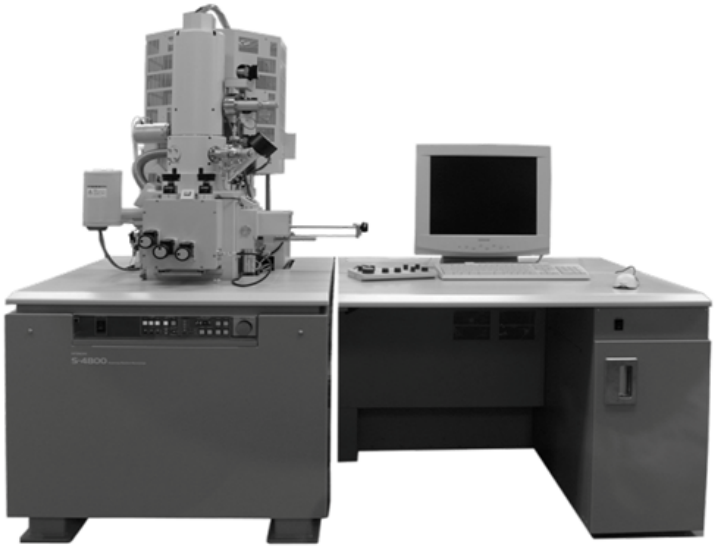


Figura 2.6: Microscopio SEM Hitachi - S4800.

Para los experimentos de este trabajo doctoral se han probado dos microscopios SEM, el Hitachi S3400N y el que se muestra en la Figura 2.6, que es el que finalmente se ha utilizado, el Hitachi S4800. Éste tiene una resolución máxima de 1 nm y opera a un voltaje de aceleración que varía entre 0,5 y 30 kV. Su capacidad de magnificación alcanza los 800.000X. Destaca la movilidad del porta muestras, el cual puede rotar 360ºy tiene una capacidad de inclinación de entre -5ºy 70º. Aunque no estaba operativo en el ejemplar al que se ha tenido acceso para la realización del presente trabajo, tiene posibilidad de STEM, un tipo de microscopio TEM explicado en el siguiente apartado.

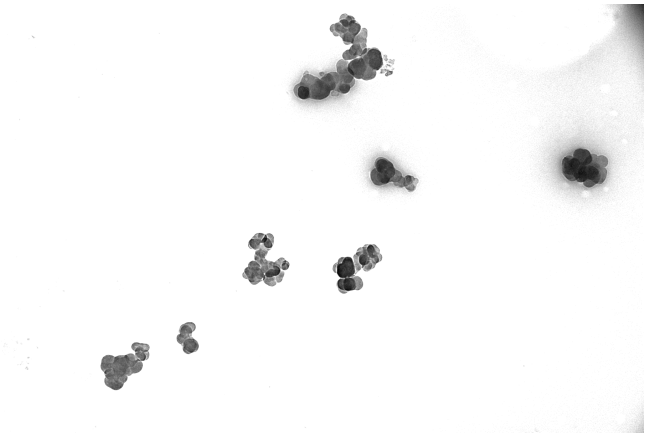


Figura 2.7: Agregados de *carbon black* capturados con un microscopio TEM.

Una variante de los microscopios SEM es el ESEM (Microscopio Electrónico de Barrido Ambiental), el cual permite observar muestras húmedas evitando dañar la muestra durante su preparación. Con esta técnica se pueden analizar muestras incluso dentro de un líquido [BTB+05].

Figura 2.8: Microscopio TEM Philips - EM208S.

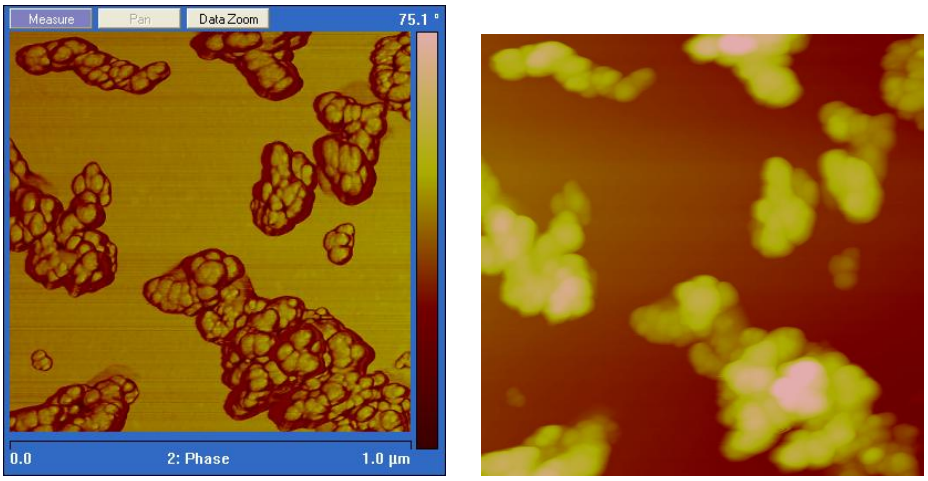
2.2.1.2 **TEM**

Los microscopios TEM (Microscopio Electrónico de Transmisión) emiten un haz de electrones hacia la muestra. Una parte de estos electrones se pierden porque rebotan o son absorbidos por el objeto de estudio. El resto, atraviesan la muestra y gracias al sistema de registro se captura la imagen aumentada. Los electrones son dirigidos y enfocados mediante lentes magnéticas en el vacío, ya que las moléculas del aire pueden desviarlos. La Figura 2.7 muestra varios agregados de *carbon black* capturados con un microscopio TEM. La forma de operación normal, llamada campo claro o *brightfield* presenta los objetos oscuros y el fondo claro, mientras que la técnica opuesta, llamada campo oscuro o *darkfield* funciona a la inversa. La diferencia se encuentra en que en el modo de campo oscuro se bloquean los electrones que inciden directamente sobre la muestra, para que la imagen se forme sólo con los electrones dispersados.

En la Figura 2.8 podemos ver el microscopio TEM que hemos utilizado para los experimentos, el Philips EM208S. Su rango de voltaje se encuentra

entre 20 y 120 kV y normalmente se trabaja entre 80 y 100 kV.

El STEM (Microscopio Electrónico de Transmisión de Barrido) es un tipo de microscopio TEM, que concentra el haz de electrones en un área pequeña, que a su vez es escaneada. Requiere un voltaje de trabajo menor que el TEM, lo que permite analizar muestras biológicas sensibles [SAZL11].



(a) **(b)**

Figura 2.9: Agregados de *carbon black* capturado con un microscopio AFM: a) imagen de fase y b) imagen de alturas.

2.2.1.3 **AFM**

Los microscopios AFM (Microscopio de Fuerza Atómica) utilizan un sensor mecánico con una punta fina que va rastreando la superficie. Al mismo tiempo, un láser apunta a este sensor, registrando las fuerzas que se dan entre la punta y la superficie. Éstas incluyen las fuerzas por contacto mecánico, fuerzas de van der Waals, fuerzas magnéticas, fuerzas electrostáticas, enlaces químicos, etc. Dispone de 3 modos de operación básicos: modo de contacto, modo de contacto intermitente o *tapping* y modo de no-contacto. Una ventaja que posee es que las muestras no requieren una preparación previa especial y puede analizar muestras introducidas en líquidos.

En la Figura 2.9 se muestra diferente información de los mismos agregados: a) imagen de fase y b) imagen de alturas. La imagen de fase muestra el cambio de fase de la onda sinusoidal aplicada a la punta. Esto refleja las diferencias de adhesión en la superficie de la muestra y se obtiene en modo de contacto intermitente. La imagen de alturas o imagen topográfica se obtiene midiendo

el movimiento vertical del tubo de barrido. Ésta, se puede adquirir tanto en modo de contacto, como en modo de contacto intermitente.

En la Figura 2.10 se muestra el microscopio utilizado, el Nanoscope IIIa de Digital Instruments. Éste puede operar en los modos de contacto y contacto intermitente. Además, lleva acoplado un microscopio óptico, gracias al cual se facilita el trabajo de situar la punta en una zona determinada de la muestra. En el modo de contacto se puede alcanzar la resolución atómica.

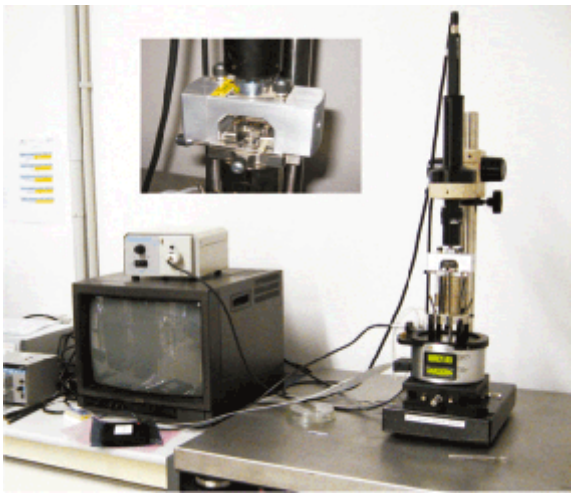


Figura 2.10: Microscopio AFM Digital Instruments - Nanoscope IIIa.

2.2.2 Estándar ASTM

El estándar ASTM (American Society for Testing and Materials) para el tratamiento de imágenes TEM para la caracterización del *carbon black* [Ame07] establece que para una correcta caracterización de los agregados de carbono hay que realizar sobre las imágenes operaciones de eliminación de ruido/fondo y *binarización* utilizando un cierto umbral. Posteriormente, y sobre la imagen *binarizada*, se habrán de realizar operaciones de erosión y dilatación además de medir el área, el perímetro y un mínimo de 16 diámetros de Feret distanciados angularmente en 11,25° [Ame02]. El umbral puede ser calculado automáticamente, manualmente o hasta se puede mantener uno preestablecido si el grosor de la película de carbono es constante y no se alteran otras condiciones de la imagen. Para el tamaño mínimo de imagen recomendado de 640x480 hay que descartar los agregados de menos de 50 píxeles de área atribuyéndolo al ruido. Para suavizar el ruido asociado a los bordes del agregado sin alterar la

información real, se puede realizar una erosión y una dilatación de un píxel de profundidad y una o dos pasadas. Por último, los agregados que tocan los límites de la imagen deben ser también descartados por estar incompletos.

2.2.3 Eliminación de ruido

El ruido se define como una perturbación indeseada en una señal. Aplicado al área de las imágenes podría definirse como un píxel que tomará valores diferentes a sus vecinos en relación al tono de gris. Por desgracia, las imágenes de microscopios electrónicos contienen mucho ruido. Incluso al utilizar aparatos de gran calidad, el ratio señal/ruido es muy bajo. Mediante un escaneo SEM lento de 20 segundos se obtienen imágenes con una relación señal/ruido de 20:1, mientras que para alcanzar un ratio de 120:1 hace unos años era necesario un tiempo de adquisición de 10 minutos [VC07]. En este último caso, la calidad es más que aceptable y además puede ser mejorada por medio del método *Blind Deconvolution* en caso de ser guardadas como imágenes TIFF de 16-bits. No obstante, no es normal disponer de tanto tiempo para la captura de las imágenes por lo que se trabaja con imágenes con alto contenido de ruido que es necesario eliminar.

Para minimizar al máximo posible el ruido se utilizarán diferentes tipos de filtros, teniendo en cuenta siempre el tipo de ruido que hay que eliminar. Existen diferentes tipos de ruido indeseado, que se pueden clasificar en:

- **Gaussiano**: produce pequeñas variaciones en la imagen. Tiene su origen en diferencias de ganancias del sensor, ruido en la digitalización, etc.
- **Impulsional**: (denominado sal y pimienta) los puntos de ruido toman valores límite, bien muy altos (blanco) o muy bajo (negro).
- **Frecuencial**: la imagen resultante puede ser la suma de la imagen ideal más una interferencia.

A la hora de eliminar las distorsiones creadas por el factor ruido existen numerosas alternativas. Los filtros pueden clasificarse en temporales y espaciales. Se emplearán los del segundo tipo, debido a que los primeros trabajan con conjuntos de imágenes tomadas en diferentes instantes de tiempo. Los espaciales lineales se fundamentan en la convolución entre la imagen original y una máscara para reducir el ruido. El principal inconveniente de estas técnicas es el enturbiamiento que se produce en la imagen, provocando el difuminado

de los bordes. El filtro gaussiano y el de la media son los más destacados en esta categoría. El filtro gaussiano, como era de esperar, es especialmente útil en la reducción de ruido gaussiano. El valor de cada nuevo punto es el resultado de promediar con distintos pesos los valores vecinos a ambos lados de dicho punto. Presenta el problema del difuminado de los bordes, pero no es tan acusado como el caso de la media simple. Con este último, dada una imagen $f(i, j)$, se genera una nueva imagen $g(i, j)$ cuya intensidad para cada píxel es calculada promediando los valores de intensidad de los píxeles $f(i, j)$ incluidos en un entorno de vecindad predefinido [PdlC07].

Asimismo, cabe destacar el filtro lineal de Wiener [Wei49]. Éste, procesa una señal de entrada $x(n)$ a la que aplica un filtro lineal de respuesta impulsional $h(n)$ con el que se obtiene una señal de salida $s(n)$. Se considera que la diferencia entre la salida $s(n)$ y la señal deseada $d(n)$ es el error de estimación $e(n)$ que se pretende minimizar en términos del MSE (Mínimo Error Cuadrático). Este filtro es además uno de los más importantes para eliminar ruido en señales de voz [CBHD06].

Por otro lado, existe el algoritmo *anisotropic diffusion*, propuesto por Perona y Malik [PM90], que ha sido utilizado con éxito para la eliminación de ruido en imágenes a escala nanométrica [FH01]. Este método emplea ecuaciones diferenciales parciales que difuminan la imagen de una manera no uniforme, al contrario que los flujos de calor, que difuminan la imagen isotrópicamente, es decir, en todas las direcciones [Sap01]. Este método de Perona y Malik, a diferencia de los filtros tradicionales de paso bajo mantiene los bordes [RS06], algo necesario para poder identificar los agregados posteriormente. La ecuación de difusión anisotrópica requiere la definición de una función de parada ante un borde, del inglés *edge-stopping* que está relacionada con el error de la normal. Perona y Malik utilizaron la función de Lorentz y posteriormente se han propuesto alternativas para esta función. Una buena opción es la función robusta de *Tukey's biweight* basada en el error de la normal, con la que se han conseguido bordes más nítidos y una mejor parada automática de la difusión [BSMH98]. Otra forma de ver el *anisotropic diffusion* es como un procedimiento robusto que estima una imagen suavizada dividiendo en *parches* la imagen original que contiene ruido [BSMH98]. Sobre una imagen suavizada por este procedimiento los bordes pueden ser identificados con gran éxito gracias a la interpretación estadística robusta. En el marco de la estimación robusta, los límites entre las regiones tienen valores atípicos, es decir, son numéricamente distantes del resto de los datos. Además, añadiendo restricciones sobre la organización de los bordes a los resultados de la ecuación de Perona y Malik se

consigue una mejora cualitativa en la continuidad de los bordes [BR96].

2.2.4 *Binarización*

Una vez disponemos de una imagen con el menor ruido posible, procedemos a su *binarización*. Esta operación consiste en transformar una imagen en color, o escala de grises en nuestro caso, en una imagen en la que cada píxel solo tiene dos valores posibles.

Existen dos alternativas para llevar a cabo esta tarea, el establecimiento de un umbral para discernir entre el fondo y los objetos, y la detección de los bordes. El primer enfoque tiene la ventaja de que es más robusto frente al ruido y el segundo, de que, en buenas condiciones, es un método más preciso.

2.2.4.1 **Umbral óptimo de la escala de grises**

Para discernir entre el fondo de la imagen y los elementos a analizar, en este caso los agregados de negro de humo, se puede definir un valor de la escala de grises a partir del cual los píxeles serán considerados parte de los elementos.

Las numerosas técnicas existentes para encontrar el umbral óptimo de la escala de grises se pueden clasificar en seis categorías dependiendo de las características de la imagen en las que se basen: forma del histograma, *clustering* del espacio, entropía del histograma, atributos de la imagen, información espacial y características locales [SS04].

Forma del Histograma

Los métodos basados en la forma del histograma analizan los picos, valles y curvaturas del histograma suavizado. El método de Rosenfeld consiste en analizar las concavidades del histograma comparándolas con su envoltura convexa [RdlT83]. Al calcular la envoltura convexa del histograma, los puntos cóncavos más profundos se convierten en candidatos para ser el valor de umbral. Para elegir entre diferentes concavidades se pueden utilizar atributos del objeto. Otras variaciones de este método tratan el problema de la búsqueda del valle [WR77b, WR77a, HOS87, SG92]. Más recientemente, se ha mejorado este método teniendo en cuenta la envoltura exponencial del histograma [Wha91].

Por otro lado está el método del pico y el valle de Sezan, que realiza un análisis de los picos aplicando una convolución con una matriz de suavizado y diferenciación [Sez90]. De forma similar, mediante el análisis del nivel múltiple del PMF (enlace peptídico de la masa de una huella), se interpretan sus

huellas, que son el recorrido de sus cambios de signo y extremos en la escala [Car87]. En otro estudio, usando una transformada de *wavelet* dual discreta consiguen un análisis multiresolución del histograma [Oli94]. Por último, se encuentra la técnica de modelado de la forma, que utiliza una función simple de aproximación al PMF formada por una función de dos pasos [RYS95]. De esta manera se minimiza la suma de los cuadrados entre una función de dos niveles y el histograma, encontrando la solución con una búsqueda iterativa. Asimismo, existen generalizaciones de la idea de la aproximación a una forma. Usando el análisis de espectro de Prony, se puede aproximar el espectro como la densidad espectral de señales exponenciales complejas [CL98]. Un modelo similar se basa en el espectro de potencia del histograma [GP98].

Clustering del espacio

En métodos de *clustering* se crean dos *clusteres*, uno para el fondo y otro para los objetos. Estos *clusteres* pueden ser modelados como la mezcla de dos funciones de Gauss. El procedimiento iterativo de Riddler fue uno de los primeros métodos iterativos basado en la mezcla de dos funciones de Gauss [RC78]. En él, en la iteración n se establece un nuevo umbral U_n usando las medias del fondo y los objetos, terminando las iteraciones cuando $|T_n - T_{n+1}|$ sea lo suficientemente pequeño.

El método Otsu calcula el umbral óptimo minimizando la suma ponderada de la varianza dentro de cada una de las dos clases (fondo y objetos), o lo que es lo mismo, que busca la maximización de la dispersión entre clases [Nob79]. Es uno de los métodos más referenciados y da sus mejores resultados cuando el número de píxeles de cada una de las dos clases es parecido. Tiene algunas limitaciones [LP90] y se propone el método Otsu de dos dimensiones [LP90] que se comporta mejor frente al ruido utilizando la correlación espacial de cada píxel con sus píxeles vecinos. Otra técnica es la búsqueda del error mínimo, con la que se plantean [Llo85] dos funciones de densidad de Gauss de igual varianza y por medio de una búsqueda iterativa minimiza el error total de clasificación. Por último está el *fuzzy clustering* [JBR97], que asigna los píxeles a un *cluster* dependiendo de las diferencias con las medias de cada clase. El umbral se establece como el punto de corte entre las funciones de pertenencia a las clases.

Entropía del histograma

Las técnicas basadas en la entropía son algoritmos que utilizan la entropía del fondo, la de los objetos y la entropía cruzada entre la imagen original y una *binarizada*. Se distinguen 3 métodos: entropía simple, cruzada y aproximada.

El primero de ellos [KSW85] considera que el fondo y los objetos son dos señales diferentes y cuando la suma de las entropías de las clases alcanza el máximo, se toma el umbral como óptimo. El enfoque cruzado [LL93] y [LT98] busca la minimización de la distancia teórica. Distancia de Kullback-Leibler de las distribuciones de la imagen real y la reconstruida, siendo requisito necesario el que la imagen reconstruida tenga la misma media de intensidad que la real tanto para los fondos como para los dos grupos de objetos. El método de la entropía aproximada [Sha94] considera los miembros difusos como un signo de la fuerza con la que un píxel pertenece al fondo o a los objetos. De hecho, cuanto más lejos esté un valor de la escala de grises de un umbral inicial (cuanto más adentrado en su región), su potencial para pertenecer a una clase específica aumenta.

Atributos de la imagen

Los procedimientos basados en los atributos de los objetos utilizan medidas de similitud entre la imagen en escala de grises y la *binarizada*, como similitud aproximada o *fuzzy* de la forma, coincidencia del borde, etc. Se clasifican en conservar el momento, coincidencia con los bordes, similitud aproximada, estabilidad topológica, máxima información y mejora de la compactación aproximada. En el primer grupo se encuentra el método Tsai [Tsa85], que considera la imagen real como una imagen emborronada de una imagen binaria ideal. El umbral se establece de forma que los tres primeros momentos sean iguales en la imagen de grises y en la binaria. Este algoritmo ha sido reformulado [CT93] usando redes neuronales. El caso de buscar la coincidencia con los bordes [HS88] es un método iterativo que compara los bordes obtenidos de la imagen real con los bordes de la imagen binaria. Para las dos imágenes se utiliza el operador de Sobel [SF68] para encontrar los bordes. El umbral definitivo será el que maximice la coincidencia de los bordes, penalizando los bordes no correspondidos tanto en un sentido como en el otro. No obstante, con esta técnica se han encontrado algunos problemas [VR95]. La similitud aproximada fue planteada inicialmente por Murthy y Pal [MP90b], y posteriormente Huang y Wang propusieron un índice de difuminado para medir la distancia entre la imagen en escala de grises y la binaria [HW95]. Posteriormente, Ramar *et ál.* evaluaron diferentes medidas aproximadas para buscar un umbral, como índice lineal, índice cuadrático, medida logarítmica de la entropía y medida exponencial de la entropía, concluyendo que el índice lineal es el que da mejores resultados [RAS+00].

Los expertos en el uso de los microscopios ajustan subjetivamente el umbral [Rus87] en un punto en el que los bordes y las formas de los objetos se

estabilizan. La aproximación de la estabilidad topológica [PA96] busca un umbral que se estabiliza cuando los objetos alcanzan su tamaño correcto. Esto se consigue con una función de tamaño-umbral que es definida como el número de objetos que contienen un número mínimo de píxeles. El umbral se establece en la mayor asíntota horizontal de la función de tamaño-umbral. El ruido desaparece rápidamente y la parte constante de la función es el rango en el que el fondo y los objetos son fácilmente diferenciables. Se elige el valor en la mitad de la mayor asíntota horizontal de la función tamaño-umbral como umbral óptimo. El método de la máxima información [LL98] define el problema del umbral como el cambio en la incertidumbre de que un píxel pertenezca a una clase u otra. La presentación de información del fondo y de los objetos reduce la incertidumbre, que es medida para elegir el umbral que la minimice. La mejora de la compactación aproximada [Tsa85] toma el perímetro y el área como funciones del umbral. El umbral óptimo maximiza la compactibilidad de los objetos segmentados. También se ha utilizado este método para la detección de pequeños objetos [Fer00] maximizando la distancia de Kolmogorov-Smirnov entre los histogramas del fondo y el de los objetos.

Información espacial

Los métodos espaciales utilizan procedimientos de distribución de probabilidades y de correlación entre píxeles. Utilizan la distribución de los valores de gris así como la dependencia de un píxel en su entorno. Se distinguen los métodos por matriz de coexistencia o *cooccurrence*, entropía de grado alto, conjuntos aleatorios y partición aproximada 2D. El uso de las matrices de coexistencia para encontrar un umbral fue propuesto por Chanda y Majumder [CM88] y en una línea parecida N.R. Pal y S.K. Pal [PP89] observaron que dos imágenes con histogramas idénticos pueden tener entropías de diferente grado por su estructura espacial. La entropía de grado alto [Abu89] tiene en cuenta la entropía conjunta de dos variables aleatorias relacionadas, concretamente, el valor de gris de un píxel y la media del valor de gris de una región de píxeles vecinos centrada en dicho píxel. En otro estudio [LGC97] se utiliza también la proyección lineal de Fisher de un histograma 2D. Por otro lado [BLNdL95] explotan la correlación espacial de los píxeles usando la entropía de bloques yuxtapuestos. La metodología de los conjuntos aleatorios considera que cada valor de umbral proporciona un conjunto de objetos binarios con diferente distancia de Chamfer [Bor86]. En la partición aproximada 2D se combinan [CC99] las ideas de la entropía aproximada, las del histograma 2D de los valores de los píxeles y el histograma de las medias de cada región de 3x3 píxeles. Brink tiene en cuenta también la entropía espacial que se refleja

indirectamente en las estadísticas de coexistencia [Bri89] y [Bri95].

Características locales

Los métodos locales adaptan el umbral de cada píxel teniendo en cuenta a los cercanos que le rodean. Existen 4 técnicas: contraste local, rodear al centro, ajuste a una superficie y método Kriging. La técnica del contraste local compara el valor de un píxel con la media en su vecindario. Esta región recomendada [WR83] de tamaño 15x15 está centrada en dicho píxel y su tamaño se elige para que aproximadamente tenga el tamaño de los objetos. Si el píxel es significativamente más oscuro que la media, se considera objeto y si no, fondo. Los métodos basados en esquemas que rodean al centro [GPS77] fueron mejorados por [PSS86], consistiendo en medir el contraste local entre 5 matrices de 3x3 píxeles. La matriz central es la encargada de capturar la parte delantera y las otras 4, en posición diagonal a la central, el fondo. Otra forma de buscar un umbral es el ajuste a una superficie [YB89]. En este método se combinan bordes con información de nivel de gris. La superficie umbral es construida por interpolación utilizando varias funciones para calcular bordes. Utilizando una red neuronal[1] de Hopfield se plantea el paradigma de la superficie activa [SI97]. El último método local es el de Kriging, con el que se realizan dos pasos. En el primero se buscan dos umbrales globales [KSW85] con los que cada píxel queda clasificado en una de las dos clases posibles y los píxeles cuyo valor se encuentre entre estos dos umbrales son los que se decidirán en el segundo paso. Este segundo paso se realiza usando la covarianza local de los indicadores de clase y una regresión lineal de Kriging en una región de 3 píxeles de radio (28 píxeles).

2.2.4.2 Detección de bordes

Una alternativa a los métodos de binarización basados en el establecimiento de un umbral (apartado 2.2.4.1) para la localización de regiones, es la detección de bordes. Esta técnica busca cambios en la escala de grises para determinar los píxeles que son bordes. El objetivo es encontrar los bordes del objeto real sin permitir que el ruido los enmascare. Éstos constituyen una transición de oscuro a claro o viceversa, y son capturados no como un cambio brusco de intensidad

[1] Las redes neuronales artificiales (del inglés, *Artificial Neural Networks (ANN)*), son un modelo matemático inspirado en la forma en la que funciona el sistema nervioso de los animales. Estas redes están compuestas por un grupo de neuronas artificiales interconectadas entre sí. En la mayoría de los casos, este tipo de redes son un sistema adaptativo que cambia de estructura en función de la información que fluye por la red en la fase de aprendizaje.

sino como una transición gradual como resultado del muestreo [PdlC07]. La primera derivada es cero en todas las regiones de intensidad constante y tiene un valor constante en toda la transición de intensidad. La segunda derivada, por otro lado, es cero en todos los puntos, excepto en el comienzo y el final de una transición de intensidad. Por tanto, un cambio de intensidad se manifiesta como un cambio brusco en la primera derivada y presenta un paso por cero, es decir, se produce un cambio de signo en su valor en la segunda derivada. Este cambio de signo es el denominado *zero-crossing*. Así, en base a la primera y segunda derivada, han surgido numerosos métodos para el reconocimiento de bordes.

Operadores primera derivada

Entre los operadores primera derivada, se encuentra el gradiente de una imagen, que se define como:

$$\boldsymbol{G}\left[f\left(x,y\right)\right] = \left[\begin{array}{c} Gx \\ Gy \end{array}\right] = \left[\begin{array}{c} \frac{\partial}{\partial x} f\left(x,y\right) \\ \frac{\partial}{\partial y} f\left(x,y\right) \end{array}\right] \tag{2.4}$$

donde $f(x,y)$ es la imagen, (x,y) es el punto para el que se está calculando el gradiente $\boldsymbol{G}$, que es un vector perpendicular al borde. $\boldsymbol{G}$ apunta en la dirección de variación máxima de la imagen en el punto en cuestión. La magnitud se calcula según la ecuación:

$$|\boldsymbol{G}| = \sqrt{G_x^2 + \ G_y^2} \tag{2.5}$$

La dirección viene dada por la ecuación:

$$(x,y) = tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{2.6}$$

Normalmente, la magnitud del gradiente se aproxima con valores absolutos:

$$|\boldsymbol{G}| \approx |G_x| + \ |G_y| \tag{2.7}$$

ya que el valor de la magnitud del gradiente no es tan importante como la relación entre diferentes valores. Para calcular las derivadas se utilizan las diferencias de primer orden entre dos píxeles adyacentes, esto es:

$$G_x = \frac{f\,(x+\Delta x) - \ f(x-\Delta x)}{2\Delta x} \qquad G_y = \frac{f\,(y+\Delta y) - \ f(y-\Delta y)}{2\Delta y} \tag{2.8}$$

No obstante, los operadores más difundidos que implementan el concepto de derivada en un punto, Sobel [SF68], Prewitt [Pre70] y Roberts [Rob65], realizan una convolución en la imagen con máscaras de 3x3. Es decir, consideran una vecindad de 3x3 centrada en el píxel del que se quiere calcular su gradiente. En la Figura 2.11 se muestran las máscaras de Sobel (b) y (c) para calcular G_x y G_y respectivamente, para el punto z_5 de la región (a).

$$\underset{(a)}{\begin{bmatrix} z_1 & z_2 & z_3 \\ z_4 & z_5 & z_6 \\ z_7 & z_8 & z_9 \end{bmatrix}} \quad \underset{(b)}{\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}} \quad \underset{(c)}{\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}}$$

$$\underset{(d)}{\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}} \quad \underset{(e)}{\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}}$$

Figura 2.11: Máscaras de convolución para calcular con la región a) el punto z_5: b) máscara de Sobel usada para obtener G_x, c) máscara de Sobel usada para obtener G_y, d) máscara de Prewitt usada para obtener G_x y e) máscara de Prewitt usada para obtener G_y.

De este modo, las derivadas basadas en los operadores de Sobel se expresan de la siguiente forma:

$$\begin{aligned} G_x &= (z_3 + \ 2z_6 + \ z_9) - \ (z_1 + 2z_4 + \ z_7) \\ G_y &= (z_7 + \ 2z_8 + \ z_9) - \ (z_1 + 2z_2 + \ z_3) \end{aligned} \tag{2.9}$$

Para el operador Prewitt se utilizan las matrices d) y e) de la Figura 2.11 de forma análoga al operador de Sobel. Por otro lado, el operador de Roberts, que trabaja muy bien en imágenes binarias, utiliza dos diagonales perpendiculares:

$$\begin{aligned} D_1 &= f(x,y) - f(x-1,y-1) \\ D_2 &= f(x,y-1) - f(x-1,y) \end{aligned} \tag{2.10}$$

y tiene dos formas posibles:

$$\begin{aligned} R &= \sqrt{D_1^2 + D_2^2} \\ R &= |D_1| + |D_2| \end{aligned} \tag{2.11}$$

Una vez calculado el gradiente de una imagen por uno de los métodos expuestos, se establece un valor U no negativo de umbral de gradiente para descartar los píxeles que no lo alcancen, es decir:

$$g(x,y) = \begin{cases} 1 & si \quad \mathbf{G}[f(x,y)] > U \\ 0 & si \quad \mathbf{G}[f(x,y)] \leq U \end{cases} \tag{2.12}$$

En la Figura 2.12 se muestra cómo al aplicar los diferentes operadores se obtienen los bordes de las imágenes. El gradiente se puede calcular tanto en horizontal (X), como en vertical (Y). En la figura se muestra las dos.

También cabe mencionar las máscaras de Kirsch [Kir71], que son 8 máscaras que se aplican por cada punto y se selecciona la que obtenga mayor resultado, las máscaras de Robinson [Rob77] se usan de forma similar a las de Kirsch y sólo cambia la primera de ellas. Las máscaras de Frei-Chen [FC77] son 9 y el proceso de proyección es similar al proceso de convolución en el sentido de que ambos superponen la máscara en la imagen, multiplican términos coincidentes y suman los resultados. Otro algoritmo importante es el de Canny [Can86], que una vez obtenido el gradiente, realiza sobre él una supresión no máxima y una histéresis del umbral.

Si la imagen contiene ruido, estos operadores pueden no dar un buen resultado. Esto se puede evitar expandiendo las matrices con las que realizar la convolución. Tamaños habituales son $7x7$, $9x9$ y $11x11$.

Operadores segunda derivada

No son tan utilizados como los operadores primera derivada, pero destacan el operador Laplaciana, Laplaciana de la Gaussiana y diferencia de Gaussianas [Lin93]. El operador Laplaciana se define como:

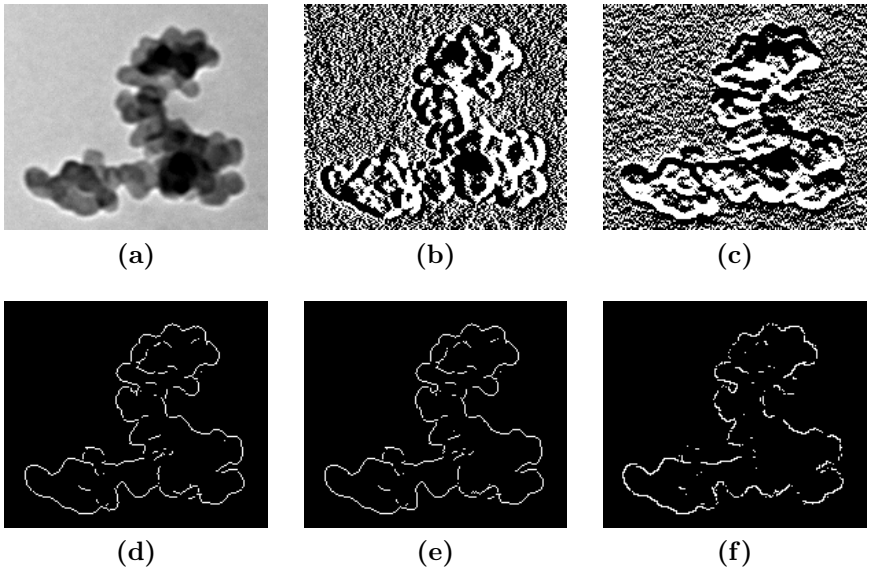


Figura 2.12: Aplicación de operadores primera derivada sobre la imagen a) agregado *carbon black*: b) magnitud del gradiente en X, c) magnitud del gradiente en Y, d) operador Sobel, e) operador Prewitt y f) operador Roberts.

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{2.13}$$

y se puede implementar de forma digital como:

$$\nabla^2 f = 4z_5 - (z_2 + z_4 + z_6 + z_8) \tag{2.14}$$

El operador Laplaciana de la Gaussiana (LG) viene dado por:

$$\nabla^2 G(x,y) = K\left(2 - \frac{x^2+y^2}{\sigma^2}\right) e^{-(x^2+y^2)/2\sigma^2} \tag{2.15}$$

donde

$$G(x,y) = -\frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2+y^2)/2\sigma^2} \tag{2.16}$$

y K es una constante de escalado que se establece para determinar el rango de valores de la LG.

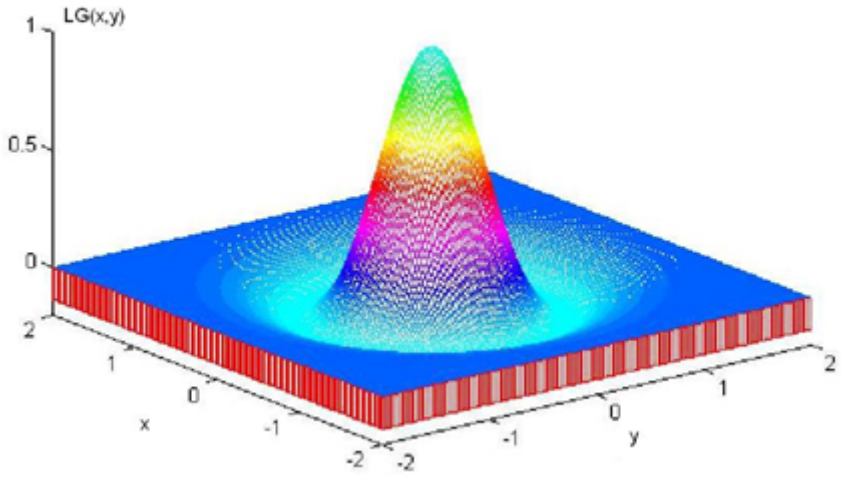


Figura 2.13: Representación del operador LG obtenido de la ecuación (2.16).

2.2.5 Segmentación

La segmentación es el proceso de delimitación de regiones dentro de una imagen. Es el paso en el que se interpreta la imagen *binarizada*. En el caso de trabajar con una imagen *binarizada* mediante un proceso de detección de bordes, como se ha expuesto en el apartado 2.2.4.2 es necesario que el perímetro quede cerrado, por esto, para imágenes con un alto nivel de ruido es más apropiada la *binarización* mediante un umbral, como se ha descrito en el apartado 2.2.4.1. Así, con el fin de optimizar las regiones se aplican diferentes operaciones morfológicas de suavizado o relleno de huecos.

2.2.5.1 Operaciones morfológicas

Las operaciones morfológicas simplifican las imágenes y preservan las formas principales de los objetos. Se utilizan principalmente en imágenes binarias y los fundamentos matemáticos fueron concebidos desde el punto de vista de la posición, que es la base del tratamiento binario, antes que desde la intensidad [PdlC07]. Las operaciones morfológicas básicas son dilatación, erosión, apertura y cierre. Otras más complejas son transformaciones homotópicas, extracción del esqueleto, adelgazamiento, ensanchado y *convex hull*.

La dilatación puede ser considerada como una transformación que cambia todos los píxeles del fondo que son vecinos del objeto. La erosión es dual de la

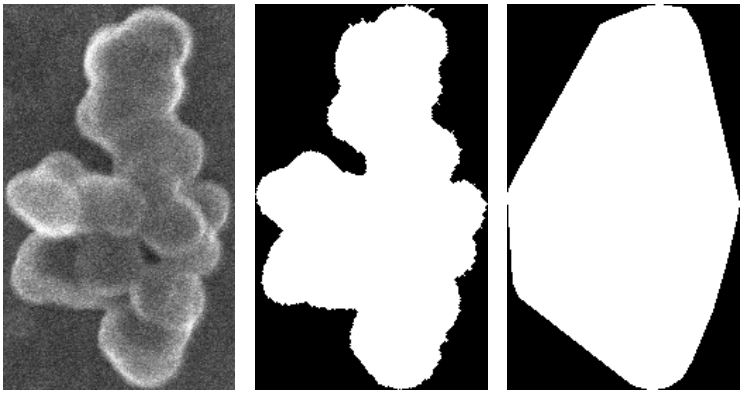


Figura 2.14: Agregado de *carbon black*: a) imagen original, b) imagen binaria y c) *convex hull*.

dilatación y se utiliza para simplificar la estructura de los objetos desgastando o hasta eliminando objetos muy pequeños que pueden ser ruido. La apertura consiste en una erosión seguida de una dilatación. Su operación dual es el cierre y consiste en realizar estas dos mismas operaciones en el orden contrario. El cierre conecta objetos que están próximos entre sí, rellena pequeños huecos y suaviza el contorno del objeto rellenando los pequeños valles mientras que la apertura produce el efecto contrario [Low91].

Una transformación homotópica es una transformación morfológica [Ser83] que no cambia la relación de contigüidad entre las regiones y los huecos de la imagen. Esta relación se expresa mediante el árbol homotópico: su raíz corresponde al fondo de la imagen, el primer nivel de ramas corresponde a los objetos (regiones), el segundo nivel se corresponde con los huecos dentro de los objetos y así sucesivamente. El esqueleto [GD88, Ser83] es la forma de la región a base del adelgazamiento de esta y por su importancia en este trabajo se expone de forma más extensa en el apartado 2.2.5.2. El *convex hull* o envoltura convexa es la forma convexa más pequeña que incluye a la región. En la Figura 2.14 se muestra a) el agregado original, b) el agregado *binarizado* y c) el *convex hull* del agregado.

2.2.5.2 **Esqueletización**

Como se ha mencionado en el apartado 2.1.5.2 la esqueletización es un proceso usado para la caracterización del *carbon black* [HMSH93, MG99]. De la misma forma, también es conocida su utilidad en campos tan dispares como el análisis en tiempo real del movimiento humano [FL98], el reconocimiento de caracteres

[BK12] o a la hora de buscar similitudes entre objetos [SM11].

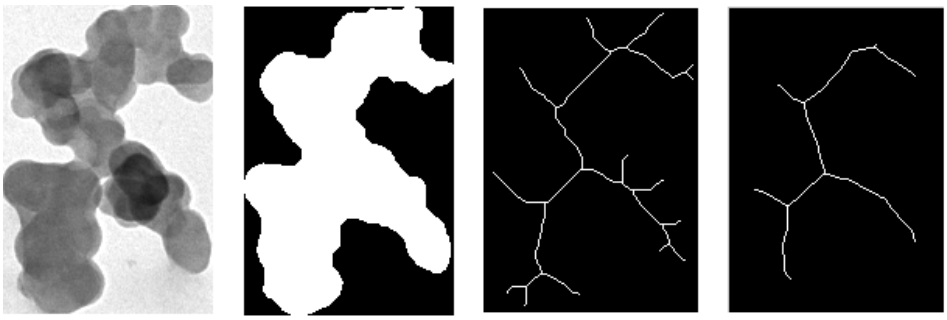


Figura 2.15: Agregado de *carbon black*: a) imagen original, b) imagen binaria, c) esqueletización tradicional y d)esqueletización simplificada.

La esqueletización [Pav80, LLS92] de una imagen digital es una técnica que se aplica sobre una región de una imagen que ha sido previamente binarizada. En concreto, consiste en eliminar sucesivamente los píxeles del borde de la región. Esta erosión se realiza analizando los píxeles vecinos. Un píxel se elimina mientras no divida la región en varias y continúa hasta que sólo quedan líneas de un píxel de grosor. Gracias a esta técnica, se puede cuantificar el número de ramas que tiene una región y su longitud [HMSH93, MG99].

En la Figura 2.15 se muestra un agregado capturado con un microscopio electrónico de transmisión, el agregado *binarizado*, y dos esqueletizaciones, siendo la segunda la implementación de Nicholas Howe [How07] basada en el trabajo de Alex Telea [TVW02, RT02], con la que se consigue un esqueleto más simplificado y robusto frente a perímetros irregulares.

2.3 Extracción de características de las imágenes

Para una correcta obtención de características geométricas –tal y como el área o diámetros de Feret– será necesario procesar las imágenes que los microscopios proporcionan. Estas características geométricas servirán más adelante para clasificar los diferentes agregados y estimar una serie de valores aproximados y estrechamente relacionados con la geometría como son el volumen de una partícula y el número de partículas por agregado. Las técnicas de procesado involucrarán etapas de eliminación de ruido, refinado y delimitación de contornos.

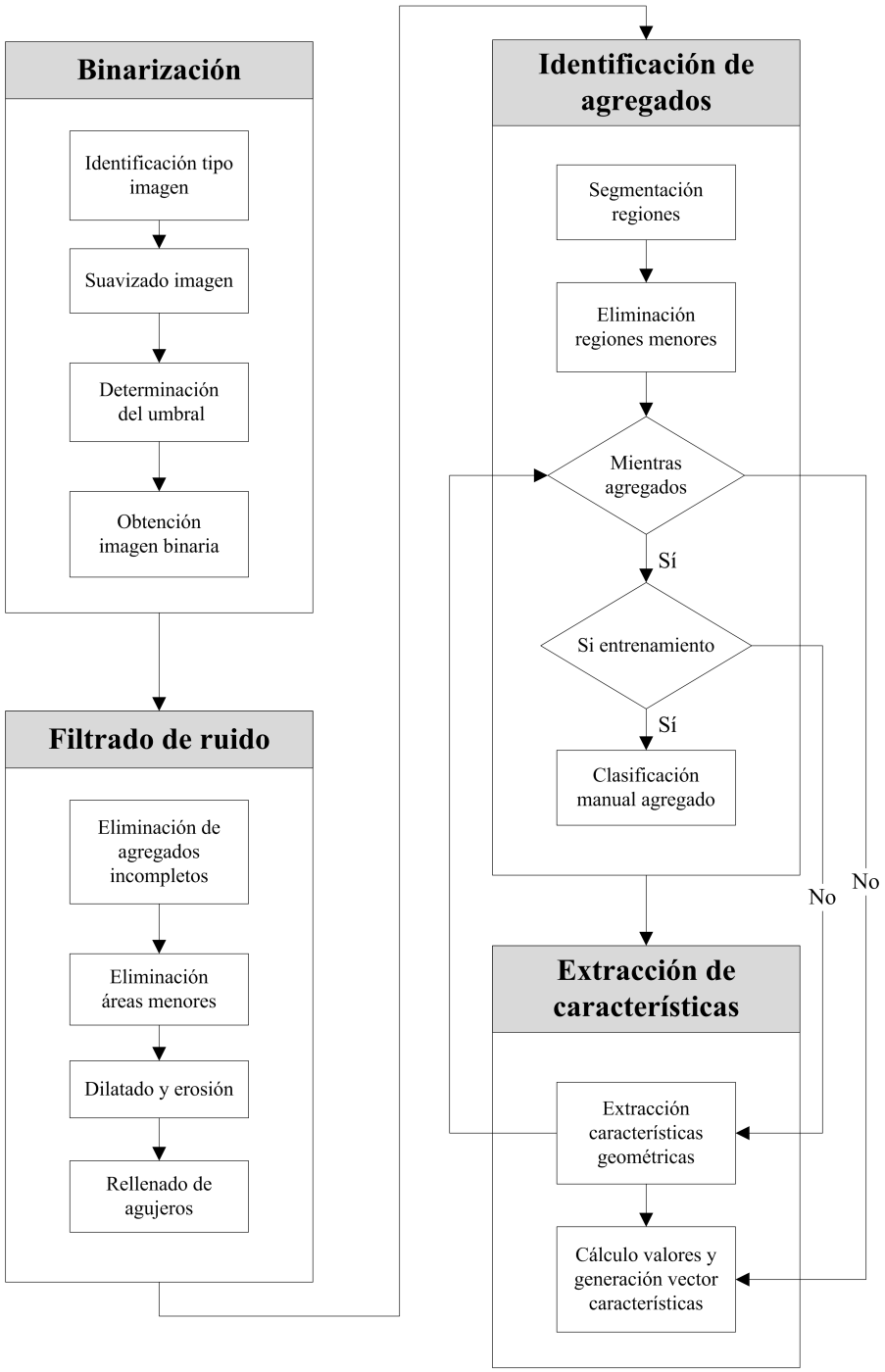


Figura 2.16: Algoritmo creado para la extracción de características de una imagen.

Las imágenes 2D con las que se ha trabajado han sido capturadas fundamentalmente vía microscopios SEM y TEM. Ambos se basan en el uso de un haz de electrones para la captura de imágenes, pero cada uno de ellos permite el estudio de diferentes características de una muestra. El SEM (Microscopio Electrónico de Barrido) provee información sobre morfología y características de la superficie, mientras que con el TEM (Microscopio Electrónico de Transmisión), podemos observar la estructura interna y detalles ultraestructurales.

Inicialmente, capturamos también agregados de *carbon black* mediante un microscopio AFM con el objetivo de aprovechar la información de profundidad que este proporciona, sin embargo se encontraron grandes dificultades a la hora de segmentar este tipo de imágenes. Según el estándar ASTM sobre la metodología a seguir para la caracterización del *carbon black* a partir de imágenes de microscopio [Ame07], hay ciertos pasos que es necesario realizar que se detallan en el apartado 2.2.2. En la Figura 2.16 se muestra el algoritmo desarrollado teniendo en cuenta este estándar.

Específicamente, el objetivo de este tratamiento consiste en segmentar los agregados para extraer de ellos varias características geométricas. No sólo las básicas, tales como el área o el perímetro, son consideradas, sino también más complejas como la *absorción* o el *perímetro fractal*. Una vez obtenidas se evaluará su relevancia y se mediante algoritmos de *machine learning* se determinará la morfología de los agregados no clasificados.

El algoritmo está formado por cuatro fases diferentes. En primer lugar, en la fase de *binarización*, apartado 2.3.1, se obtiene una imagen en blanco y negro. En segundo lugar, en el apartado 2.3.2, filtramos el ruido de la imagen. En tercer lugar, en el apartado 2.3.3, se identifican todos los agregados. Por último, en el apartado 2.3.4 se delimitan los agregados y se extraen sus características geométricas.

2.3.1 Fase I: *Binarización*

Para realizar una *binarización* de la imagen, esto es, pasar de una imagen en escala de grises a una imagen cuyos píxeles solo puedan tomar el valor negro o blanco, seguimos los siguientes pasos:

1. **Identificación del tipo de imagen:** En base a la ruta de la imagen se determina si la imagen es SEM o TEM para el ajuste de los parámetros de procesado. Los agregados son más claros que el fondo en las imágenes SEM, y más oscuros que el fondo en las imágenes TEM. Con el objetivo

de identificar correctamente los agregados invertimos el valor de la escala de grises de las imágenes TEM, con lo que tendremos también agregados claros sobre fondo oscuro.

2. **Suavizado de la imagen:** las imágenes de microscopios electrónicos contienen una cantidad de ruido que dificulta una detección de bordes precisa [LY08]. De hecho, obtener una imagen con un ratio de señal-ruido de 120:1 requería 10 minutos hace unos años [VC07]. Por tanto, la herramienta que hemos desarrollado permite aplicar diferentes filtros: (i) el suavizado gaussiano [PdlC07], un operador de convolución bidimensional usado para eliminar detalle y ruido; (ii) el filtro de difusión anisotrópica [PM90], un filtro que elimina el ruido mientras mantiene los bordes; y (iii) Wiener [Wei49], un filtro adaptativo también para la eliminación del ruido.

3. **Determinación del umbral:** Para obtener una imagen binaria de calidad es necesario encontrar un umbral adecuado. Este umbral o *threshold*, es un valor de la escala de grises que se utiliza para dividir los píxeles en blancos o negros. Este valor es crucial para conseguir una imagen binaria adecuada de la que se puedan segmentar elementos [Ots75]. Por tanto, es conveniente calcular este valor por un método fiable, como el de Otsu [Ots75], un método no supervisado, sin parámetros de configuración para calcular automáticamente un valor de umbral. Ya que en algunos casos puede no resultar el umbral óptimo, nuestra herramienta permite también su establecimiento manual o la adición de una constante al valor determinado por Otsu.

4. **Obtención de la imagen binaria:** Una vez que se ha determinado el umbral, generamos la imagen binaria considerando para cada píxel, que pertenece al fondo si su valor de la escala de grises no alcanza el umbral impuesto, y por el contrario, si lo supera, es considerado como parte de un agregado.

2.3.2 Fase II: Filtrado de ruido

Aunque se realice un suavizado en la fase de *binarización* para eliminar el ruido, pueden quedar elementos no deseados en la imagen. Estos elementos pueden ser fácilmente confundidos con los agregados que estamos intentando segmentar, por esto, es imprescindible que los eliminemos. La fase de filtrado de ruido está compuesta por los siguientes pasos:

1. **Eliminación de áreas menores:** en las imágenes de microscopios electrónicos con las que trabajamos es fácil encontrarse con regiones que realmente pertenecen al fondo. Éstas poseen una intensidad de la escala de grises similar a los agregados, por lo que no son clasificadas apropiadamente y aparecen como puntos blancos en la imagen binaria. Por tanto, eliminamos estos agregados irreales antes de llegar a la fase de reconocimiento para evitar falsas detecciones.

2. **Dilatado y erosión:** para mejorar la calidad de los bordes, creamos un elemento estructural con forma de disco que es utilizado para realizar una operación de cierre que consiste en dilatar y posteriormente erosionar una imagen. Como resultado, los bordes quedan suavizados. En la Figura 2.17 se muestra un disco con radio de 3 píxeles, que es en realidad una matriz de vecindad que se aplica a cada píxel para calcular su nuevo valor. Así al dilatar una imagen, un píxel pasará a tener valor 1, es decir blanco, con que uno de los *vecinos* de la matriz sea 1 en la imagen original. De manera análoga, la erosión establece a 0 los píxeles que tengan algún vecino con valor 0. De esta forma se corrigen píxeles que tienen un valor por encima del umbral de *binarización* por ser ruido o por estar en el límite de dicho umbral [Rus07].

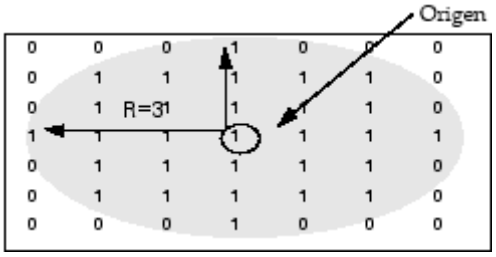


Figura 2.17: Elemento estructural con forma de disco para aplicar operaciones de dilatado y erosión a las imágenes. Los unos son los vecinos a tener en cuenta para calcular el nuevo valor del origen.

3. **Eliminación de agregados incompletos:** según el estándar ASTM [Ame07], los agregados que tocan el borde de la imagen están incompletos y no pueden ser utilizados como parte del estudio. Por esto, llegado este paso, las regiones que contienen por lo menos un píxel en el perímetro de la imagen son eliminadas.

4. **Rellenado de agujeros:** esta tarea detecta áreas que realmente son parte de un agregado, pero que por ser oscuras, en la fase de *binariza-*

ción se han identificado como parte del fondo. En otras palabras, identificamos los puntos negros que se encuentran dentro de áreas blancas cerradas y corregimos su valor.

2.3.3 Fase III: Identificación de agregados

Una vez terminada la segunda fase, obtenemos una imagen filtrada donde se puede distinguir perfectamente a los agregados. Además, hemos detectado y eliminado áreas incompletas. Sin embargo, la imagen todavía no está lista para que se le extraigan características geométricas, ya que en una sola imagen hay varios agregados. Para ello, con el objetivo de identificarlos, ejecutamos una tercera fase, que se compone de los siguientes pasos:

1. **Delimitación de agregados o segmentación:** analizamos la imagen y recogemos información de todas las regiones o elementos existentes. Para asegurarnos de que cada elemento se corresponde con un agregado, realizamos una última verificación. Una partícula de *carbon black* se define como un pequeño componente esferoidal que forma parte de un agregado y no se encuentran por separado [DBW93]. Así, establecemos que el tamaño mínimo de un agregado es el que contiene dos partículas. Los grados de *carbon black* utilizados en esta investigación tienen un diámetro de 50 nm o superior, por lo que se puede afirmar que un agregado no tendrá un área inferior a $2 \cdot \pi \cdot 25^2 = 3.927\ nm^2$. A partir de este umbral descartamos las áreas que sean menores.

   Una vez que las áreas que superan el umbral son identificadas, recortamos la imagen binaria en trozos, cada uno de ellos conteniendo un agregado. Ya que al *binarizar* la dimensión de la imagen no cambia, utilizamos los mismos valores para recortar la imagen original. Finalmente, guardamos en una misma imagen el agregado en versión original y *binarizada*, como se puede apreciar en la Figura 2.18.

2. **Clasificación de agregados:** para poder entrenar y validar los algoritmos de aprendizaje automático descritos en la sección 3.2, realizamos un etiquetado manual de los agregados. Este etiquetado es necesario para entrenar a los algoritmos de aprendizaje automático y que posteriormente sean capaces de clasificar nuevos agregados. Como ya se ha mostrado en la Figura 2.2 al explicar los métodos directos de caracterización en el apartado 2.1.5.2, hay cuatro categorías en las que se pueden clasifi-

car: circular, esferoidal, lineal y ramificada. Además, esta clasificación es también necesaria para la evaluación de la relevancia de los atributos.

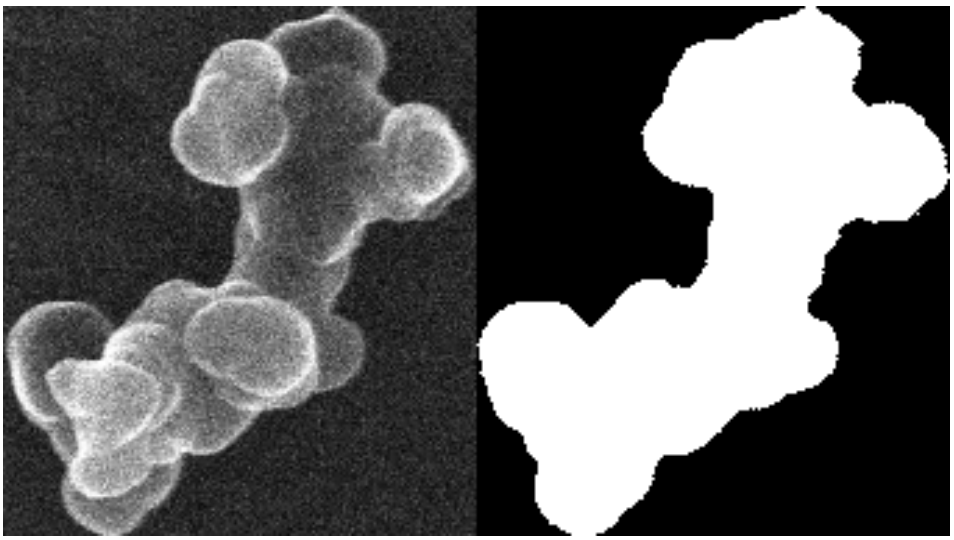


Figura 2.18: Imagen SEM original y *binarizada*

2.3.4 Fase IV: Extracción de características

Utilizando la salida de la fase anterior, extraemos 45 características geométricas de las imágenes. Estas las dividimos en 5 grupos. Comenzamos con las 4 características consideradas imprescindibles por el estándar ASTM [Ame07] para el tratamiento de imágenes de microscopio de *carbon black*. En el segundo grupo situamos los 6 parámetros que el mismo estándar estima a partir del área y el perímetro. Posteriormente, hemos utilizado 2 cálculos parciales necesarios para determinar dos de los parámetros de distribución incluidos en el informe que establece el estándar ASTM [Ame07], pero que por sí solos no se utilizan en la literatura. A continuación, hemos incluido 20 atributos basados en la literatura para caracterizar formas, tanto para el *carbon black* como para objetos en general.

Por último, describimos los 13 atributos que aportamos al estado de la técnica. En la sección 2.6 se evalúa la relevancia de todos los atributos con respecto a la clase morfológica de los agregados.

En teoría, para describir la forma de un objeto se deben utilizar métricas adimensionales [Rus07], es decir, que no tienen dimensión. Así, la relación de

aspecto, que divide el ancho entre el alto y de esta forma, las dos medidas, que están en la misma unidad se anulan.

Por ejemplo, dos objetos de diferente tamaño pero con la misma forma tendrán distinta área, lo que puede parecer irrelevante para describir la forma. En cambio, con la circularidad, definida como $4 \cdot \pi \cdot área/perímetro^2$ se anulan las dimensiones nm^2/nm^2 y en este caso los dos objetos darán el mismo valor de circularidad. Sin embargo, en el caso de los agregados de *carbon black*, un agregado circular tendrá un área pequeña, ya que es muy difícil que los que pertenecen a esta categoría tengan muchas partículas. Así, los agregados más grandes tienen más probabilidades de ser ramificados.

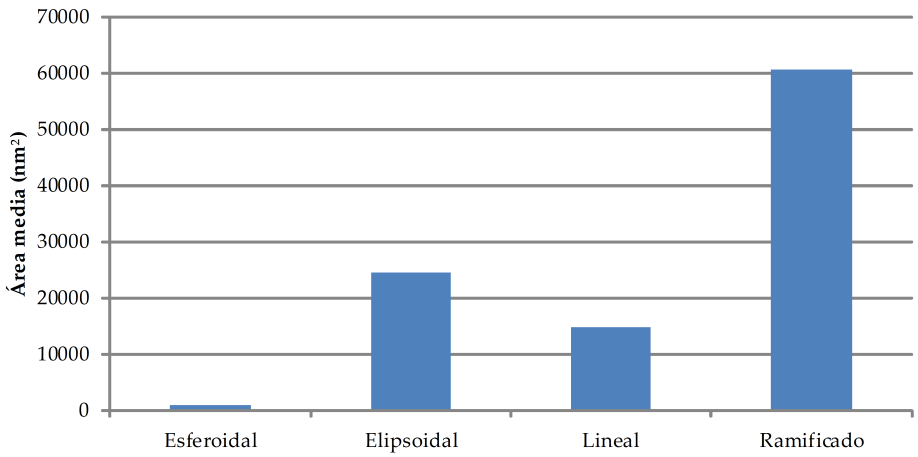


Figura 2.19: Área media de los agregados por cada tipo de morfología para el conjunto SEM.

En la Figura 2.19 queda demostrado para el conjunto de datos SEM inicial 2.4.1, cómo la media del área para los agregados esferoidales, 756 nm^2, es notoriamente inferior al resto de las clases. Del mismo modo, para los agregados ramificados, la diferencia también es destacable, siendo su media de 60.643 nm^2. En cambio, para diferenciar entre elipsoidales y lineales no es una característica tan diferenciadora, 24.470 nm^2 y 14.781 nm^2 respectivamente. Aun así, se han querido diferenciar los parámetros dependientes del tamaño por medio de un asterisco (*).

Mediciones ASTM

Las 4 características que se describen a continuación son consideradas imprescindibles por el estándar ASTM [Ame07] para el tratamiento de imágenes de

microscopio de *carbon black*.

- **área (*):** definida como la extensión de la proyección bidimensional del agregado. Al tratarse de nano-partículas trabajamos en nm^2.

- **perímetro (*):** es la longitud de la línea que rodea el área, se mide en nm.

- **feretMáximo (*)**: diámetro máximo de Feret en nm. Un diámetro de Feret se define como la distancia entre dos tangentes paralelas que tocan al agregado en lados apuestos. Siguiendo la norma ASTM se comprueban 16 diámetros con una separación de 11,25º[Ame02]. El máximo Feret se utiliza también para orientar el agregado, lo que influye en ciertas características, que se describirán posteriormente: *centroideX, centroideY, sectoresVacíos, cambiosDerivada, porcentajeSegmentos, porcentajeCortesSegmentos, áreaÁreaBoundingBox, mediaÁreaÁreaConvexa, desviaciónÁreaÁreaConvexa*, *mediaÁreaÁreaTriángulo* y por último, en *desviaciónÁreaÁreaTriángulo*. De esta forma se obtienen valores más similares para un agregado independientemente de la diferencia rotacional que pueda existir en el eje perpendicular al plano de la proyección que es la imagen.

- **feretMínimo (*):** diámetro mínimo de Feret en nm.

Parámetros ASTM estimados

En este segundo grupo se encuentran 6 características definidas por el estándar ASTM [Ame07] que se estiman a partir del área y el perímetro.

- **diámetroEquivalente (*):** diámetro equivalente de un círculo con la misma área que el agregado en nm. Calculado como $\sqrt{4 \cdot área/\pi})$.

- **factorDeAgregación:** definido como $13,092 \cdot (perímetro^2/área)^{-0,92}$. Se establece un límite inferior a $0,4$, es decir, que si es menor de $0,4$ se determina que es $0,4$.

- **diámetroPartícula (*):** estimación del diámetro de partícula en nm calculado según la relación: $factorDeAgregación \cdot \pi \cdot área/perímetro$. Se trata de un parámetro adimensional.

- **volumenAgregado (*):** el volumen de un agregado, al igual que el de una partícula, se estima utilizando principios estereológicos[1]. Se da en nm^3 y se calcula como $(8/3) \cdot área^2/perímetro$.

- **volumenPartícula (*):** el volumen de una partícula, en nm^3, se estima como $\pi \cdot diámetroPartícula^3/6$.

- **númeroPartículasAgregado:** es la estimación del número de partículas existentes en el agregado. Esta característica se calcula por la relación $volumenAgregado/volumenPartícula$. Se trata de un parámetro adimensional.

Cálculos parciales para parámetros de distribución ASTM

En el tercer conjunto se sitúan dos características que no han sido utilizadas en la literatura en sí mismas, pero son necesarias para realizar cálculos de distribución en el informe que establece el estándar ASTM [Ame07].

- **tamañoPartícula (*):** se calcula como $númeroPartículasAgregado \cdot diámetroPartícula$, en nm.

- **volumenEsfera (*):** volumen de la esfera con diámetro del área equivalente del agregado, en nm^3, calculado mediante la ecuación $4/3 \cdot \pi \cdot (tamañoMedioAgregado\ /2)^3$. Donde *tamañoMedioAgregado*, el diámetro del área equivalente del agregado, queda determinado por $2 \cdot \sqrt{área/\pi}$.

Atributos para caracterizar formas basados en la literatura

En el cuarto grupo se encuentran 20 atributos que han sido utilizados con anterioridad para caracterizar formas, tanto para el *carbon black* como para objetos en general.

- **perímetroÁrea (*):** determinado por la relación entre el *perímetro* y el *área* [SXM08]. Las unidades de este parámetro son el inverso del nm.

- **feretMínimoMáximo:** relación entre los diámetros de Feret menor y mayor, *feretMínimo / feretMáximo*. El atributo F_3 empleado por Hess *et ál.* [HMU73] y Herd *et ál.* [HMH92] ha sido invertido para que se encuentre entre 0 y 1. Es un parámetro adimensional.

[1] La estereología es una técnica que permite calcular información tridimensional a partir de información bidimensional, es decir, trata de la interpretación tridimensional de objetos observados bidimensionalmente.

- **feret90 (*):** diámetro de Feret perpendicular al Feret mayor en *nm* [Ama03].
- **relaciónFeret90FeretMáximo:** relación entre el diámetro de Feret perpendicular al Feret mayor y el Feret mayor, es decir, *feret90 / feretMáximo*. De esta forma, el parámetro longitud-anchura utilizado por Herd *et ál.* [HMH92, HMSH93], queda normalizado, con un rango entre 0 y 1, y utilizando la longitud mayor real del agregado independientemente de la orientación con la que ha sido capturado. Esta relación es similar a *feretMínimoMáximo*, sin embargo, teniendo en cuenta que el *feretMínimo* no es siempre ortogonal al *feretMáximo*, este ratio aporta información adicional. Es un parámetro adimensional.
- **perímetroFractal (*):** calculado como $-2 \cdot \log(área)/log(perímetro)$ en base a la relación dimensión-perímetro-área $perímetro \sim área^{D_p/2}$ [HMH92, GO91, MG99], explicada en el apartado 2.1.5.2.
- **factorOclusión (*):** dado por la diferencia entre $volumenEsfera$ y $volumenAgregado$ [HMH92, HMSH93]. Su dimensión es nm^3.
- **absorción:** calculado como la relación entre el $volumenAgregado$ y el *factorOclusión*. Es el inverso del parámetro empleado por Herd *et ál.* [HMH92]. Es un parámetro adimensional.
- **circularidad (*):** indica la similitud del agregado a un círculo. Su valor es 1 para un círculo y tiende a 0 para un segmento cuanto más fino y largo sea. Se calcula como $4 \cdot \pi \cdot área/perímetro^2$ [Pra07] y se da en nm^2. También conocido como *formfactor* [Rus07] o *Fcircle* [MG99]. Otros autores [Jäh02] definen la circularidad como $perímetro^2/área$.
- **áreaConvexa (*):** área del menor polígono convexo que envuelve al agregado en nm^2 [PI97]. Ha sido explicado en el apartado 2.2.5.1.
- **relaciónÁreaÁreaConvexa:** relación entre el área del agregado y el área del polígono convexo, $área/áreaConvexa$, también conocido como *solidity* [PI97, Rus07]. Es un parámetro adimensional con el que a menor ratio tenemos un agregado con mayor irregularidad.
- **áreaÁreaBoundingBox:** relación entre el área del agregado y el área del *Bounding Box*[1] [Rus07]. Éste no se rota para ajustarse a él. También es conocido como *extent* y es un parámetro adimensional.

[1]El *Bounding Box* es un término que describe la caja menor que envuelve a un objeto.

- **perímetroConvexoPerímetro:** relación entre el perímetro del polígono convexo y el perímetro original, conocido como *convexity* [Rus07]. Es un parámetro adimensional.

- **númeroRamas:** contabiliza el número de ramas secundarias, es decir, todas menos la principal. Es un parámetro adimensional que requiere la esqueletización del agregado [HMSH93, MG99]. En el estado de la técnica sobre el tratamiento de imagen, se dedica el apartado 2.2.5.2, a describir los diferentes métodos de esqueletización. Finalmente, en los experimentos presentados, se ha utilizado la esqueletización simplificada [How07].

- **mediaRamas (*):** media de la longitud de las ramas del esqueleto en píxeles [MG99]. Para el cálculo del esqueleto redimensionamos la imagen del agregado a 200 píxeles de altura he indicamos su valor en píxeles para representar su tamaño proporcional respecto al esqueleto. Lo mismo sucede con el cálculo de *desviaciónRamas, medianaRamas, mediaRecortadaRamas, mediaCentralRamas*.

- **centroideX:** relación entre la coordenada horizontal del centro de masa de la región binaria y la longitud horizontal del *Bounding Box* [Rus07]. En nuestro caso, si *centroideX* es mayor que 0,5 se cambia por 1 – *centroideX* para que si volteamos un agregado horizontalmente obtengamos el mismo valor. Se trata de un parámetro adimensional.

- **centroideY:** característica análoga al *centroideX*, pero con la coordenada vertical del centro de masa de la región binaria [Rus07]. Es un parámetro adimensional.

- **ejeMayorElipse (*):** longitud en *nm* del eje mayor de la elipse que tiene el mismo segundo momento central que la región [Med71, PI97].

- **ejeMenorElipse (*):** longitud en *nm* del eje menor de la elipse que tiene el mismo segundo momento central normalizado que la región [Med71, PI97].

- **relaciónEjes:** relación entre el eje menor de la elipse y el mayor, calculado como $ejeMenorElipse/ejeMayorElipse$ [PI97]. Es el inverso de una característica definida por Medalia [Med71]. Se trata de un parámetro adimensional.

Así, todos los lados de la caja tocan al objeto que envuelven. En nuestro caso, por tratarse de una región bidimensional, se trata de un rectángulo que envuelve al agregado.

- **excentricidad:** del inglés *eccentricity*, mide la excentricidad de la elipse que tiene el mismo segundo momento central normalizado que la región [Ama03]. La excentricidad es el ratio de la distancia entre los focos de la elipse (esto es, los dos puntos desde los cuales la distancia a todos los puntos de la elipse es constante) y la longitud de su eje mayor, el $ejeMayorElipse$. Tiene valores entre 0 y 1, siendo 0 cuando la región es un círculo y 1 cuando la región es un segmento. Es un concepto similar a la circularidad, pero es adimensional y proporciona información diferente, lo que puede corroborarse en la sección 2.6 al evaluar su relevancia.

Atributos propios

Por último, describimos los 13 atributos que aportamos al estado de la técnica.

- **mediaÁreaÁreaConvexa:** para calcular este parámetro se divide la imagen en 8 sectores triangulares siendo el *centroide* de la región un punto común a todos ellos y formados de tal manera que ocupan toda la superficie del *Bounding Box*. En cada sector, se calcula la relación entre el área y el área convexa y finalmente se calcula la media de estas 8 relaciones. Es un parámetro adimensional.

- **desviaciónÁreaÁreaConvexa:** desviación estándar de las 8 relaciones anteriores. Se trata de un parámetro adimensional.

- **mediaÁreaÁreaTriángulo:** similar a *mediaÁreaÁreaConvexa*, pero en este caso la relación es entre el área ocupada por la región y el área del triángulo. Es un parámetro adimensional.

- **desviaciónÁreaÁreaTriángulo:** similar a *desviaciónÁreaÁreaConvexa*, pero en este caso la relación es entre el área ocupada por la región y el área del triángulo. Se trata de un parámetro adimensional.

- **sectoresVacíos:** número de sectores triangulares que no contienen agregado. Aunque pueda parecer que por tocar el *centroide* deberían cortar al agregado esto no es necesario. Se trata de un parámetro adimensional.

- **cambiosDerivada:** es la relación entre el total de productos de la derivada segunda negativos y el total de productos calculados. Se realiza sobre el perímetro suavizado teniendo en cuenta el propio punto y los dos más cercanos. Es un parámetro adimensional.

- **porcentajeSegmentos:** número de segmentos que cortan al agregado entre el total de segmentos. Para ello se proyectan 50 segmentos aleatoriamente. Se trata de un parámetro adimensional.

- **porcentajeCortesSegmentos:** número de intersecciones de los segmentos entre el total de segmentos. Se trata de un parámetro adimensional.

- **cambioÁreaSuavizada:** indica en qué medida se ve afectada la relación entre el área y el área convexa antes y después de realizar un proceso de suavizado mediante el método de la media móvil (del inglés, *moving average*). El proceso de suavizado se realiza 10 veces teniendo en cuenta el punto actual y los 8 más cercanos, quedando la relación $(área/áreaConvexa)/(áreaSuavizada/convexÁreaSuavizada)$. Es un parámetro adimensional.

- **desviaciónRamas (*):** desviación estándar de la longitud de las ramas del esqueleto en píxeles.

- **medianaRamas (*):** mediana de la longitud de las ramas del esqueleto en píxeles. Es el valor central de las longitudes ordenadas.

- **mediaRecortadaRamas (*):** a partir de las longitudes ordenadas de las ramas del esqueleto, se elimina el tercio de los valores más altos, el tercio de los valores más bajos y se calcula la media de los que quedan, en píxeles.

- **mediaCentralRamas (*):** la media de tendencia central (en inglés *winsorizada*) es un parámetro similar a *mediaRecortadaRamas*, pero en vez de eliminar el tercio superior e inferior, se igualan al valor más próximo más alto y más bajo respectivamente, también en píxeles.

Parámetros de distribución ASTM

Por último se ha calculado un parámetro de distribución establecido por la organización ASTM [Ame07]. Éste, se determina para cada conjunto, por lo que no se utiliza ni para la evaluación de la relevancia de los atributos ni para la categorización morfológica.

- **EMSA:** el área superficial del microscopio electrónico, del inglés *Electron Microscopy Surface Area*, es el área media sin tener en cuenta la

porosidad superficial. Está inversamente relacionada con el tamaño de partícula sin considerar la porosidad. Se calcula de la siguiente forma:

$$EMSA(m^2/g) = \frac{6000}{\rho \cdot d_{sm}} \tag{2.17}$$

donde ρ es la densidad del *carbon black*, 1,8 g/cm^3 y d_{sm} es el diámetro medio de la superficie de una partícula, del inglés *particle size surface mean diameter*:

$$d_{sm}(nm) = \frac{\sum(n \cdot d_p^3)}{\sum(n \cdot d_p^2)} \tag{2.18}$$

donde n es el *númeroPartículasAgregado* y d_p es el *diámetroPartícula*.

Finalmente, generamos un vector de entrenamiento $\vec{v} = (v_1, v_2, \ldots, v_n)$ por cada agregado conteniendo todas las características expuestas, excepto el parámetro de distribución EMSA. Concretamente, cada posición v_n del vector representa una característica geométrica y se guarda con una precisión de 6 decimales. La colección de vectores forma el *dataset* o conjunto de datos de entrenamiento para el sistema de clasificación.

2.4 Conjuntos de datos

En los trabajos científicos siempre es aconsejable el utilizar conjuntos de datos públicos para que la comunidad científica sea capaz de replicar los experimentos propuestos y especialmente para demostrar los avances expuestos con respecto a publicaciones anteriores sobre los mismos datos. En el ámbito del aprendizaje automático es una práctica muy común y destaca el repositorio UC Irvine Machine Learning Repository [FA13] con conjuntos de datos muy dispares, desde características de lirios para clasificarlos en tres clases hasta información de células para detectar las cancerígenas. En cambio, en el campo de los nanomateriales no es común publicar conjuntos de datos de imágenes, sino imágenes seleccionadas, a menudo artísticas [Wer12, Nan12] con lo que no se puede llegar a formar un conjunto de datos sobre un material en concreto.

Por esto, hemos recopilado nuestros propios conjuntos de datos. Los polvos de *carbon black* analizados contienen aglomerados formados por agregados unidos entre sí por fuerzas de *Van der Waals*. Para separarlos y poder estudiar

los agregados de forma adecuada se han dispersado en una solución de cloroformo con una sonda ultrasónica modelo Bioblock Scientific Vibracell 75043 que puede observarse en la Figura 2.20.

Figura 2.20: Sonda ultrasónica Bioblock Scientific Vibracell 75043.

Posteriormente, dependiendo de la técnica de microscopía a emplear, se utilizarán soportes SEM o rejillas TEM. Una vez capturadas las imágenes con el microscopio, la herramienta creada es capaz de tratarlas automáticamente. Sin embargo, para el proceso de obtención del conjunto de datos de entrenamiento y validación es necesario prestar atención a los agregados segmentados. Así, es importante revisar los agregados en busca de duplicados. Es normal encontrarlos ya que al ir capturando imágenes, algún agregado puede encontrarse involuntariamente en dos imágenes diferentes. El problema que supone es que el conjunto de datos de validación tiene que ser diferente al de entrenamiento y si hay agregados repetidos, existe la posibilidad de que uno de ellos esté en el conjunto de entrenamiento y el otro en el conjunto de validación, sobreajustando el modelo y proporcionando una valoración irreal de él.

Cabe destacar que la tarea de etiquetado manual no es una tarea trivial, ya que hay agregados cuya forma se encuentra en un punto intermedio de los ejemplos de referencia provistos por Herd [HMH92]. Esto puede confundir al clasificador, por lo que es importante la consistencia a la hora de decidir a qué clase pertenecen.

2.4.1 SEM

Para la conferencia *DEXA (International Conference on Database and Expert Systems Applications)* creamos un conjunto de datos de 266 agregados de la

siguiente forma [LdRS+10].

Inicialmente, obtuvimos varias imágenes con tres microscopios electrónicos a diferentes escalas de magnificación. De éstas, 13 imágenes se obtuvieron con 2 microscopios SEM: el Hitachi S-3400N y el Hitachi S4800, y 11 con un microscopio TEM: el Philips EM208S. Como resultado de una evaluación previa acabamos eligiendo el segundo microscopio SEM. En la Figura 2.21 se muestra una de las imágenes que han formado este conjunto de datos. Los materiales utilizados han sido los grados de *carbon black* VULCAN XC72R y CSX691,

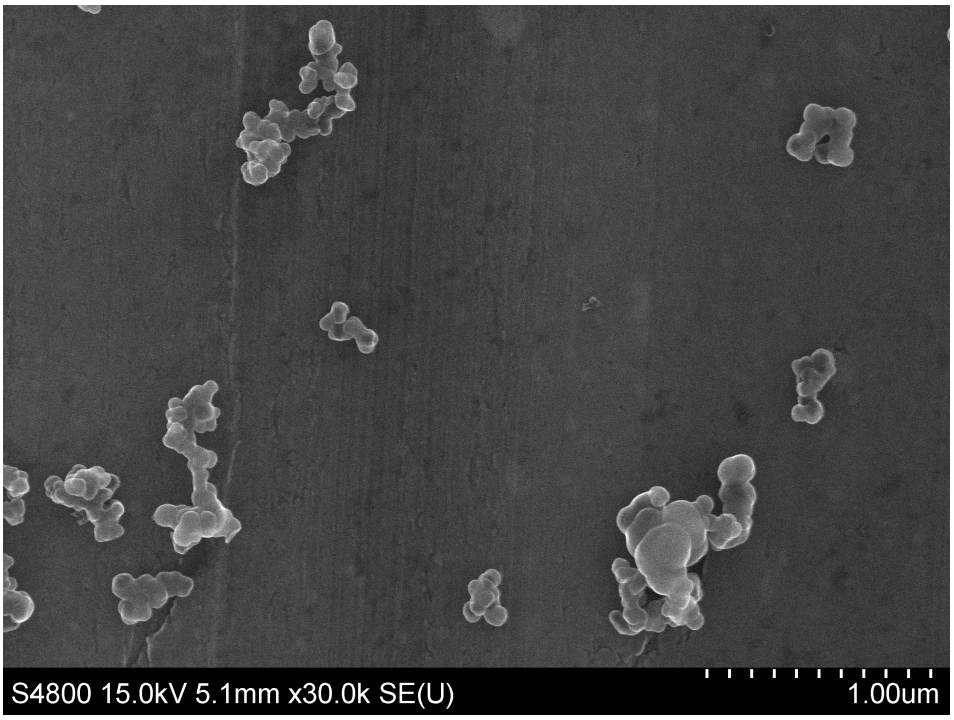


Figura 2.21: Imagen SEM de muestra.

De este modo, recogimos 144 imágenes de agregados de *carbon black* con un microscopio Hitachi S-4800 SEM. La magnificación elegida fue la de 30.000 aumentos, con una media de 3 agregados por imagen, resultando en 266 agregados correctamente segmentados que finalmente conformaron el conjunto de datos del estudio. De ellos, 9 agregados eran de tipo esferoidal, 86 del tipo elipsoidal, 51 del lineal y 120 ramificados. Posteriormente se realizó una purga adicional y un reetiquetado más sistemático para evitar la subjetividad de esta tarea, reduciendo el conjunto a 200 agregados y quedando distribuidos como 4 esferoidales, 76 elipsoidales, 33 lineales y 87 ramificados.

2.4.2 TEM

Posteriormente, creamos un conjunto de datos a partir de 198 imágenes TEM del microscopio Philips EM208S. Dichas imágenes fueron capturadas a 11.000 y 31.000 aumentos. Después de filtrar los incorrectamente segmentados y los repetidos, nos quedamos con 781 agregados, de los cuales 9 son esferoidales, 211 elipsoidales, 162 lineales y 399 ramificados. En la Figura 2.22 se muestra una de las imágenes TEM, de la cual se han extraído 5 agregados que han formado parte de este conjunto de datos. Hay que recordar que los agregados que están en contacto con el borde son descartados de acuerdo con la norma ASTM [Ame07].

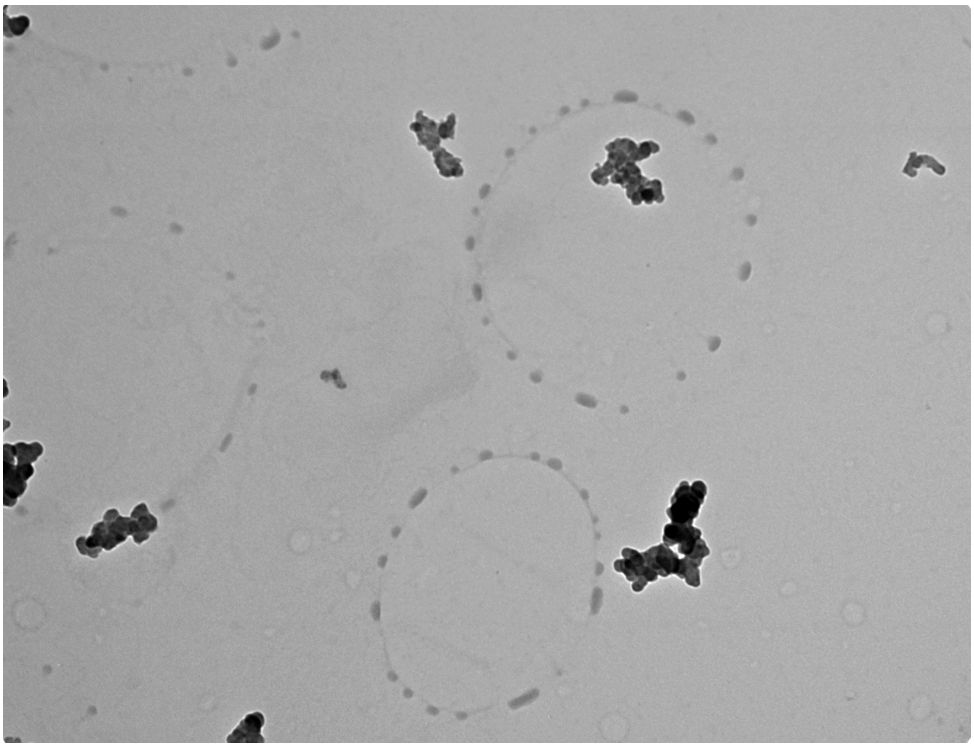


Figura 2.22: Imagen TEM de muestra.

Los materiales analizados han sido los grados de *carbon black* Vulcan XC605 y CSX691 suministrados por la empresa Cabot S.A. Para la conferencia *Modelling of Elastomeric Materials and Products* se obtuvieron además de éstas, imágenes del grado Vulcan XC72R [LdGO$^+$10]. Sin embargo, para esta tesis se han descartado las imágenes de este último grado, ya que no se han considerado aceptables por su alto contenido en residuos y por su mala iluminación.

2.4.3 Artificiales

Además, para validar la herramienta, se creó un conjunto de datos artificial con 1.788 agregados creados manualmente mediante la aplicación Paint de Microsoft®. De ellos, 399 eran esferoidales, 463 elipsoidales, 453 lineales y 473 ramificados. En la Figura 2.23 se muestra una de las imágenes que han formado este conjunto de datos. En dicha figura, todos los agregados mostrados son ramificados.

Figura 2.23: Imagen artificial de muestra.

La particularidad de este conjunto de datos frente a los otros dos es la distribución frecuencial de las clases morfológicas. Así, este conjunto de datos está balanceado, es decir, contiene un número similar de agregados por cada clase. El no tener los datos balanceados, puede suponer un problema [EHG07] y será discutido en el capítulo 3.

2.5 Selección de atributos

Cuando se trabaja con algoritmos de aprendizaje automático es frecuente intentar reducir el número de atributos con el que se trabaja. Esto tiene diversos

beneficios. El más directo es el de reducir el tiempo de extracción de las características, así como el tiempo de generación de un modelo y el tiempo de clasificación de cada atributo. En algunos casos, el tiempo de proceso requerido puede ser de vital importancia, aunque en nuestro caso priorizamos una mejor clasificación. Además de eliminar atributos, estas técnicas permiten evaluar su relevancia e identificar las características que mejor definen a un objeto a la hora de ayudar a su clasificación, con lo que se consigue un mayor conocimiento del problema en cuestión [GE03].

Otros beneficios que se pueden dar al reducir el número de atributos son el mejorar la clasificación y hacerla más robusta frente al ruido y variaciones e incluso para evitar el sobreajuste (del inglés, *overfitting*) [SIL07].

El sobreajuste se da cuando un modelo es excesivamente complejo o tiene demasiadas características en comparación con el número de instancias analizadas. El problema que supone un modelo sobreajustado, es que describe incluso los errores, lo que le hace empeorar su capacidad de predicción ante evidencias no conocidas. Es decir, está ajustado al conjunto de entrenamiento, pero no es capaz de generalizar en nuevos datos. Dicho de otro modo, un modelo se sobreajusta si es más complejo que otro que se ajusta igual de bien al conjunto de datos [Haw04]. Para evitarlo, existen diferentes técnicas, poda, parada temprana o como ya se ha mencionado, la reducción de atributos.

Por otro lado, no para evitar el *overfitting*, pero si para obtener la calidad real del algoritmo ante nuevos datos existe la técnica de la validación cruzada (del inglés, *cross validation*). En esta tesis, se ha seguido este método, que consiste en dividir el conjunto de datos en varias particiones, frecuentemente y en nuestro caso es 10. De esta forma, se entrena el modelo con todas las particiones menos una y se valida con la partición restante. Este proceso se repite tantas veces como particiones para que cada partición sea usada una vez como validación [Bis06].

2.5.1 Clasificación de los métodos

Los métodos de selección de atributos se dividen dependiendo de la forma en la que se combinan con el clasificador: filtros, encapsulados y embebidos.

- Las **técnicas de filtrado** observan sólo las propiedades intrínsecas de los datos para calcular su relevancia. Son de las más usadas por su rapidez, escalabilidad e independencia del clasificador, lo que supone menos trabajo si se quieren probar diferentes algoritmos de clasificación, pero

a la vez, al ignorar la forma en la que interactúan con el clasificador, no explotan sus diferencias en el tratamiento de la información. Existen dos enfoques, univariables y multivariables, es decir, relacionan una variable individualmente con la clase a predecir, o tienen en cuenta la influencia del resto de las variables o atributos. Las técnicas univariadas más utilizadas son χ^2 (Chi-cuadrado, del inglés, *Chi-squared*) [IMK93], distancia euclidiana, Ganancia de Información (del inglés, *Information Gain*)[Ken83] y la Relación de Ganancia de Información (del inglés, *Information Gain Ratio*) [Qui86a]. Se ha observado que el enfoque univariable es el más utilizado hasta ahora, sin embargo, los enfoques multivariables han emergido con fuerza en la última década [SIL07]. Los filtros multivariables son computacionalmente menos complejos que los encapsulados pero menos escalables que los filtros univariables. Destacan los filtros MBF(*Markov Blanket Filtering*) [KS96] y la selección rápida de atributos basada en la correlación [YL04]. Por otra parte, el filtro *ReliefF* [Kon94, RŠK03, KPH07] balancea los atributos teniendo en cuenta las interacciones entre ellos, utilizando todas las características del conjunto de entrenamiento y los K vecinos más cercanos de cada instancia para ir actualizando los pesos de cada característica.

- Las **técnicas *wrapper*** encapsulan el modelo de clasificación en la búsqueda del subconjunto óptimo de atributos. Para cada algoritmo de clasificación que se quiera probar, es necesario repetir el proceso de entrenamiento y evaluación para cada subconjunto de atributos, teniendo la ventaja de obtener el subconjunto óptimo para un conjunto de datos y algoritmo concretos. Ya que el número de subconjuntos posibles crece exponencialmente con el número de atributos, se emplean métodos heurísticos para guiar la búsqueda del subconjunto óptimo. Estos métodos pueden ser deterministas o aleatorios. Los primeros son más simples, pero corren el riesgo de sobreajustarse y de quedarse atascados en mínimos locales. Destacan los métodos de búsqueda secuencial hacia delante y hacia detrás [Kit78]. Los métodos aleatorios tienen menor riesgo de atascarse en mínimos locales pero mayor riesgo de sobreajustarse y son muy caros computacionalmente hablando. Destacan los métodos basados en algoritmos genéticos [Hol75] y el algoritmo de estimación de la distribución [ILES00].

- Por último, en las **técnicas embebidas**, la búsqueda del subconjunto óptimo se introduce en la construcción del algoritmo y puede considerarse como una búsqueda en un espacio de subconjuntos de atributos y

de hipótesis. Al igual que los métodos encapsulados, son específicos de los algoritmos de aprendizaje automático, no obstante son más rápidos. Destacan los basados en árboles de decisión, Bayes ingenuo (*Naïve Bayes*) balanceado [DHS01], vector balanceado del SVM (*Support Vector Machine*) [GWBV02, WEST03].

2.5.2 Métodos utilizados

Para la evaluación de la relevancia de los atributos hemos elegido cuatro métodos diferentes. Por un lado, hemos escogido dos técnicas de filtrado univariables para estudiar la relevancia de las características de manera individual, Chi-cuadrado y Relación de Ganancia de Información. Por otro lado, hemos seleccionado dos algoritmos multivariables para estudiar la relevancia de las características teniendo en cuenta sus interacciones entre ellas, el filtro *ReliefF* y el filtro embebido basado en el clasificador *Random Forest*. La elección de este último se debe a que por la metodología cíclica seguida en esta investigación sabemos que es el algoritmo que mejor categoriza los agregados de *carbon black*.

Chi-cuadrado

El método Chi-cuadrado, (χ^2, del inglés, *chi-squared*) [IMK93], es una técnica de filtrado univariable. Para cada atributo x_j el valor de chi-cuadrado se calcula según la ecuación (2.19):

$$Chi\text{-}cuadrado(x_j) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij} - c_{ij})^2}{c_{ij}} \tag{2.19}$$

donde n es el número de valores que puede tomar la clase (atributo a predecir), en nuestro caso 4, m es el número de valores que puede tomar el atributo x_j (los atributos continuos se discretizan[1]), a_{ij} es el número de instancias de la clase c_i que poseen el valor v_j, c_{ij} se estima mediante la ecuación (2.20):

$$c_{ij} = \frac{T_{ci} \cdot T_{vj}}{T} \tag{2.20}$$

donde T_{ci} es el número total de instancias que pertenecen a la clase c_i, T_{vj} es

[1]Discretizar consiste en transformar el valor continuo de una variable en uno discreto. Esto se consigue definiendo intervalos sobre los posibles valores que puede tomar la variable. Existen diferentes métodos, como que los intervalos sean regulares, o que contengan el mismo número de elementos para un conjunto de datos.

el número total de instancias que toman el valor v_j para el atributo analizado, y T es el número total de instancias.

Relación de Ganancia de Información

El procedimiento Relación de Ganancia de Información (del inglés, *Information Gain Ratio*) [Qui86a] es al igual que Chi-cuadrado una técnica de filtrado univariable. Se basa en la Ganancia de Información [Ken83], pero supera su problema de premiar a los atributos que poseen muchos valores diferentes [HJ02].

La ganancia de información se calcula como la diferencia de entropía antes y después de utilizar el atributo que se está estudiando para dividir el conjunto de datos según su clase. La entropía de un atributo A es una medida de la incertidumbre o desorden del atributo con respecto a la clase y se define para una clase binaria (a/b) tal y como se muestra en la ecuación (2.21):

$$Entropía(A) = -P(a) \cdot \log(P(a)) - P(b) \cdot \log(P(b)) \tag{2.21}$$

Tras partir el conjunto de datos en dos mediante un árbol de decisión, se calcula la entropía en cada una de las ramas y se balancea según el número de instancias que hay en la rama. La disminución de entropía tras la partición es la ganancia de información obtenida.

Por otra parte, para calcular la Relación de Ganancia de Información, se normaliza la Ganancia de Información dividiéndola entre el contenido de información, que se define para un atributo A en la ecuación (2.22) como:

$$Contenido\ información(A) = -\sum_{i=1}^{v} \frac{a_i + b_i}{a + b} \log_2 \left(\frac{a_i + b_i}{a + b} \right) \tag{2.22}$$

donde a_i es el número de instancias con el mismo valor Ai pertenecientes a la clase a y b_i las pertenecientes a la clase b, a es el número total de instancias pertenecientes a la primera clase y b el número de las pertenecientes a la segunda, v es el número de valores diferentes que puede tomar el atributo A.

ReliefF

Relief es un filtro multivariable, propuesto por Kira y Rendell [KR92], se basa en la idea de analizar lo discriminantes que son los valores de un atributo para las instancias cercanas. Con este fin, el método busca para cada instancia, su instancia más cercana de su misma clase, llamado el *acierto más cercano* y su

instancia más cercana de la otra clase, llamado el *fallo más cercano*.

Posteriormente, fue extendido por Konenko [Kon94] como *ReliefF*, permitiendo el análisis multiclase, analizando los fallos más cercanos a cada clase diferente. Los fallos son balanceados mediante la probabilidad de cada clase, $P(C)$. En la ecuación (2.23) se muestra cómo la función se va actualizando a medida que se recorren las diferentes instancias:

$$W[A] = W[A] - \frac{diff(A,R,H)}{m} + \sum_{C \neq clase(R)} \frac{P(C) \cdot diff(A,R,M(C))}{m} \qquad (2.23)$$

donde *diff(A,R,H)* es la diferencia existente para el atributo A entre la instancia R y la más cercana de su misma clase, H. Para atributos no numéricos o discretos la diferencia es 0 si son iguales y 1 si son diferentes. Para los atributos numéricos, se calcula la diferencia entre sus valores y se normaliza entre 0 y 1. *diff(A,R,M(C))* es la diferencia existente para el atributo A entre la instancia R y la más cercana de clase diferente. Se calculan las diferencias a cada clase diferente más cercana. Por último, m es el número de instancias analizadas.

Random Forest

El último método utilizado ha sido la evaluación de los atributos mediante un clasificador [HFH+09, BFH+11], lo que se conoce también como filtro embebido. El clasificador elegido finalmente ha sido el *Random Forest* de 1.000 árboles, que más tarde, en el capítulo 3 demuestra ser el más adecuado para categorizar este tipo de datos.

El *Random Forest* [Bre01], un conjunto de árboles de decisión construidos aleatoriamente, se describe en mayor profundidad en el apartado 3.2.4 a la hora de explicar los algoritmos de aprendizaje automático.

2.5.3 Conclusiones

Hemos visto que los métodos de selección de atributos tienen 3 objetivos: (i) mejorar la capacidad de predicción de los clasificadores, (ii) que el proceso de extracción de características y el de clasificación sean más rápidos, y (iii) llegar a entender mejor el proceso de extracción de información y sus características [GE03].

Además, hemos conocido que las técnicas de selección de atributos univariables, como Chi-cuadrado, tienen la ventaja de su simplicidad, pero al no

tener en cuenta la interacción con el resto de variables, pueden dar resultados de relevancia nulos. Así, un atributo que puede parecer prescindible, en realidad, contribuye a aumentar información al combinarse con otros atributos. Un ejemplo de esto es el conocido problema de la operación XOR al usarse como la paridad de dos bits [RB06]. A partir de una sola de las dos dimensiones es imposible conocer la clase y los métodos univariados indicarán que su relevancia es inexistente. Sin embargo, la combinación de las dos dimensiones sí que supone información suficiente para predecir la clase.

2.6 Evaluación empírica

En este apartado se presentan y discuten los resultados obtenidos con diferentes métodos para la evaluación de la relevancia de los atributos sobre los tres conjuntos de datos descritos en la sección 2.4. De cada agregado de *carbon black* se han extraído 45 características (ver apartado 2.3.4), de las cuales 13 son contribuciones de esta investigación.

En concreto, los algoritmos utilizados serán dos univariables: χ^2 o Chi-cuadrado (del inglés, *chi-squared*) y Relación de Ganancia de Información (del inglés, *Information Gain Ratio*); y dos que tienen en cuenta las dependencias entre atributos: *ReliefF* y selección de atributos mediante un filtro con el clasificador *Random Forest* [HFH+09, BFH+11] de 1.000 árboles embebido. Estos métodos determinan la importancia de un atributo en función de la clase morfológica de un agregado.

Además, se valorará la diferencia entre analizar atributos individualmente y aprovechar la interacción entre atributos. En concreto, en este apartado se busca responder a las siguientes preguntas:

- *¿Son las características aquí presentadas válidas?*
- *¿Cuáles son las características más relevantes?*
- *¿Es posible comparar las diferentes características en términos de relevancia para su uso en tareas de clasificación?*
- *¿Cómo son las características diseñadas en comparación con las existentes en la literatura?*

Para responder a las dos primeras preguntas hemos empleado dos métodos univariables: Chi-cuadrado y Relación de Ganancia de Información. Con ellos

se analizará la calidad de los atributos individualmente. Para responder a la tercera pregunta hemos utilizado dos métodos multivariables: *ReliefF* y selección de atributos mediante el clasificador *Random Forest*. Con estos dos se analizará la relevancia de cada atributo teniendo en cuenta su interacción con el resto de atributos. Por último, para responder a la cuarta pregunta se emplearán los cuatro métodos de selección de atributos, dándose a conocer la importancia de las características existentes en la literatura para la extracción de información de agregados de *carbon black* y de objetos en general.

Así, para disponer de las características morfológicas ha sido necesario realizar un procesado de las imágenes acorde a cada conjunto de datos. La manera de establecer los parámetros óptimos de *binarización* se ha realizado en tres pasos. Primero, se han llevado a cabo varias pruebas modificando interactivamente los parámetros. Una vez identificados los rangos razonables de cada uno de ellos se han *binarizado* varias imágenes con las siguientes combinaciones. Umbral de *binarización* determinado por Otsu o fijado a 0,5. Filtrado de difusión anisotrópica activado o no. Tamaño mínimo de los elementos a 3, 6, 10, 100, 500 y 1000 píxeles. Radio del disco empleado para el dilatado y erosión de 2, 4, 8, 12 y 20 píxeles. Además de este dilatado y erosión, también se añade la opción de realizarlo una segunda vez con un tamaño de disco tres veces mayor. Con esta configuración obtuvimos 240 imágenes de cada imagen original. Por último, después de seleccionar las mejores imágenes, se han terminado de ajustar manualmente los parámetros. A continuación, en la evaluación empírica de cada conjunto de datos se detallan los parámetros seleccionados.

2.6.1 Conjunto SEM

Para el tratamiento de las imágenes del conjunto SEM se ha seguido el algoritmo descrito en el apartado 2.3. En concreto, la configuración de este ha sido la siguiente. En la fase de *binarización* no se ha aplicado ningún suavizado preliminar, ya que en este conjunto empeoraba más imágenes de las que mejoraba y el umbral de *binarización* se ha ajustado a 0,4, ya que este valor resultó ser más adecuado que el obtenido mediante el método Otsu [Ots75]. En la fase de filtrado de ruido, se han eliminado los elementos que no alcanzaban los 3 píxeles de área. Además el disco empleado para el dilatado y la erosión ha sido de 2 píxeles de radio. El EMSA (Área Superficial del Microscopio Electrónico) obtenido ha sido de 55,13 m^2/g para este conjunto.

En la Tabla 2.2 se muestra el ranking obtenido según los métodos Chi-

Tabla 2.2: Relevancia de los atributos según los métodos a) Chi-cuadrado y b) Relación de Ganancia de Información sobre el conjunto de datos SEM.

(a) Chi-cuadrado

Ranking	Atributo
310,607	relaciónÁreaÁreaConvexa
242,782	áreaÁreaBoundingBox
226,835	porcentajeSegmentos
214,172	mediaÁreaÁreaConvexa
174,038	porcentajeCortesSegmentos
156,956	circularidad
156,956	absorción
156,956	númeroPartículasAgregado
145,801	factorDeAgregación
132,156	excentricidad
132,156	relaciónEjes
124,469	tamañoPartícula
109,211	perímetroConvexoPerímetro
103,372	perímetro
89,956	relaciónFeret90FeretMáximo
82,187	feretMáximo
79,397	feretMínimoMáximo
66,933	ejeMayorElipse
62,224	perímetroFractal
59,381	desviaciónÁreaÁreaConvexa
52,800	cambioÁreaSuavizada
51,890	factorOclusión
50,620	mediaÁreaÁreaTriángulo
45,867	ejeMenorElipse
44,287	feret90
43,980	áreaConvexa
41,785	feretMínimo
32,061	diámetroEquivalente
32,061	volumenEsfera
32,061	área
31,882	mediaCentralRamas
26,263	desviaciónÁreaÁreaTriángulo
24,556	sectoresVacíos

(b) Relación Ganancia Información

Ranking	Atributo
0,376	relaciónÁreaÁreaConvexa
0,325	mediaÁreaÁreaConvexa
0,324	factorDeAgregación
0,294	porcentajeCortesSegmentos
0,294	áreaÁreaBoundingBox
0,282	porcentajeSegmentos
0,255	númeroPartículasAgregado
0,255	absorción
0,255	circularidad
0,246	tamañoPartícula
0,244	perímetroConvexoPerímetro
0,221	desviaciónÁreaÁreaConvexa
0,203	excentricidad
0,203	relaciónEjes
0,202	feretMínimoMáximo
0,202	ejeMayorElipse
0,189	perímetro
0,185	perímetroFractal
0,180	relaciónFeret90FeretMáximo
0,159	ejeMenorElipse
0,157	mediaÁreaÁreaTriángulo
0,155	feret90
0,155	feretMáximo
0,150	factorOclusión
0,150	cambioÁreaSuavizada
0,147	desviaciónÁreaÁreaTriángulo
0,134	áreaConvexa
0,126	feretMínimo
0,114	mediaCentralRamas
0,091	área
0,091	diámetroEquivalente
0,091	volumenEsfera
0,090	sectoresVacíos

cuadrado y Relación de Ganancia de Información. El mejor atributo según los dos métodos univariables, *relaciónÁreaÁreaConvexa* es de hecho un parámetro comúnmente usado en la literatura [PI97, Rus07]. Según el método Chi-cuadrado se encuentra a continuación el atributo *áreaÁreaBoundingBox*, utilizado en la literatura para el tratamiento de imagen pero no en concreto para el *carbon black*.

A continuación, se encuentran en ambos métodos aunque no en el mismo orden *porcentajeSegmentos*, *mediaÁreaÁreaConvexa* y *porcentajeCortesSegmentos*, atributos diseñados en esta investigación, el primero y tercero calculados a base de trazar segmentos aleatorios. Aunque *mediaÁreaÁreaConve-*

xa obtiene una buena calificación, se basa en *relaciónÁreaÁreaConvexa*. Sin embargo, en este caso el agregado se divide en 8 sectores, y se calcula la media de este ratio. Cabe destacar también la importancia de los atributos *circularidad*, *absorción, númeroPartículasAgregado y factorDeAgregación* y que los 9 primeros atributos coinciden según los dos métodos.

En cuanto al resto de parámetros propuestos en esta investigación, no obtienen una calificación destacable y son mejor valorados por el método Relación de Ganancia de información, a excepción de *cambioÁreaSuavizada*. Destaca el que *mediaCentralRamas* sí que tiene algo de relevancia, a pesar de que el resto de parámetros relacionados con las ramas no la tengan por sí solos: *númeroRamas, mediaRamas, desviaciónRamas, medianaRamas* y *mediaRecortadaRamas*.

La excentricidad (parámetro *excentricidad*), ya se había explicado al describir las características en el apartado 2.3.4 y representa un concepto similar al atributo circularidad. Sin embargo, al estudiar su relevancia, podemos observar cómo para los dos métodos de selección de atributos univariables, *excentricidad* y *circularidad*, su relevancia es diferente. En cambio el atributo *relaciónEjes* obtiene el mismo valor que *excentricidad* por estar muy relacionados, lo que no implica que aporten la misma información, pero es necesario prestar atención especial a estas similitudes. Si se observan gráficamente los dos parámetros, muestran una tendencia similar pero invertida. Esto se debe a que ambos utilizan información de la elipse con el mismo segundo momento central que el agregado para ser calculados.

Asimismo, en el apartado 2.3.4, también se anunciaba que los parámetros *relaciónFeret90FeretMáximo* y *feretMínimoMáximo* proporcionaban información similar y así se puede confirmar según los métodos de selección de atributos, Chi-cuadrado y Relación de Ganancia de Información, en los que su valoración es muy parecida.

Se han excluido los parámetros *diámetroPartícula, volumenAgregado, mediaRecortadaRamas, volumenPartícula, cambiosDerivada, perímetroÁrea, centroideX, centroideY, númeroRamas, mediaRamas, desviaciónRamas, medianaRamas* y *mediaRecortadaRamas* ya que según Chi-cuadrado y Relación de Ganancia de Información tenían un valor nulo. Como se ha explicado en el apartado 2.5.3 esto no quiere decir que no sean relevantes para predecir la clase. Por eso, es necesario probar otros métodos para comprobar si, aunque por sí solos no aporten información, sí lo hagan en combinación con otros atributos. Esto ha sido estudiado mediante los métodos *ReliefF* y *Random Forest*, cuyos resultados se presentan en la Tabla 2.3.

Tabla 2.3: Relevancia de los atributos según los métodos a) *ReliefF* y b) *Random Forest* sobre el conjunto de datos SEM.

(a) *ReliefF*

Ranking	Atributo
0,081	porcentajeSegmentos
0,070	relaciónÁreaÁreaConvexa
0,068	áreaÁreaBoundingBox
0,060	circularidad
0,058	relaciónEjes
0,052	factorDeAgregación
0,049	feretMínimoMáximo
0,049	relaciónFeret90FeretMáximo
0,044	excentricidad
0,042	perímetroConvexoPerímetro
0,039	porcentajeCortesSegmentos
0,037	absorción
0,035	perímetroFractal
0,034	mediaÁreaÁreaTriángulo
0,030	desviaciónÁreaÁreaTriángulo
0,028	diámetroPartícula
0,026	desviaciónÁreaÁreaConvexa
0,025	mediaÁreaÁreaConvexa
0,023	ejeMenorElipse
0,022	feretMínimo
0,022	feret90
0,021	mediaRamas
0,021	númeroPartículasAgregado
0,021	mediaCentralRamas
0,020	diámetroEquivalente
0,019	ejeMayorElipse
0,019	cambioÁreaSuavizada
0,018	feretMáximo
0,017	tamañoPartícula
0,016	mediaRecortadaRamas
0,015	medianaRamas
0,013	volumenPartícula
0,013	perímetro
0,013	área
0,012	cambiosDerivada
0,010	áreaConvexa
0,010	perímetroÁrea
0,009	volumenAgregado
0,008	númeroRamas
0,007	volumenEsfera
0,007	sectoresVacíos
0,006	factorOclusión
0,004	desviaciónRamas
0,003	centroideY
0,001	centroideX

(b) *Random Forest*

Ranking	Atributo
61,423	porcentajeSegmentos
59,925	factorDeAgregación
57,303	relaciónÁreaÁreaConvexa
56,180	áreaÁreaBoundingBox
55,431	porcentajeCortesSegmentos
52,809	númeroPartículasAgregado
52,435	ejeMenorElipse
52,435	circularidad
52,060	absorción
50,936	tamañoPartícula
49,813	mediaÁreaÁreaConvexa
48,689	relaciónEjes
48,315	perímetroConvexoPerímetro
48,315	excentricidad
47,566	relaciónFeret90FeretMáximo
46,067	perímetro
45,318	cambioÁreaSuavizada
45,318	perímetroFractal
44,944	númeroRamas
44,195	factorOclusión
43,446	desviaciónÁreaÁreaTriángulo
43,446	feret90
43,446	sectoresVacíos
41,948	volumenEsfera
41,948	diámetroEquivalente
41,948	área
41,573	mediaÁreaÁreaTriángulo
41,573	ejeMayorElipse
41,199	feretMínimo
40,824	desviaciónÁreaÁreaConvexa
39,700	diámetroPartícula
39,700	volumenPartícula
38,577	feretMáximo
37,453	centroideX
36,330	volumenAgregado
35,206	áreaConvexa
35,206	perímetroÁrea
34,831	desviaciónRamas
34,831	mediaRecortadaRamas
34,082	centroideY
33,708	mediaRamas
33,333	feretMínimoMáximo
32,584	mediaCentralRamas
32,210	medianaRamas
28,839	cambiosDerivada

Así, estos dos algoritmos de selección de atributos multivariables, tienen en cuenta las interacciones entre atributos, o dicho de otra forma, su importancia dentro de un contexto. En cuanto a las similitudes con los méto-

dos Chi-cuadrado y Relación de Ganancia de Información, se encuentran los atributos mejor valorados, como: *porcentajeSegmentos, relaciónÁreaÁreaConvexa, áreaÁreaBoundingBox* y *factorDeAgregación*. En concreto, el atributo *porcentajeSegmentos*, diseñado en la presente investigación, es elegido tanto por *ReliefF* como *Random Forest* como el atributo más relevante.

En cuanto a las diferencias con respecto a los métodos univariados, el atributo *porcentajeCortesSegmentos*, altamente relacionado con *porcentajeSegmentos* pasa a obtener una peor valoración a pesar de que este último ha mejorado. Como se explicaba al describir las características en el apartado 2.3.4, para ambos descriptores se proyectan 50 segmentos aleatorios y se analizan sus intersecciones con el perímetro del agregado. Para *porcentajeSegmentos*, se contabiliza el número de segmentos que cortan al agregado entre el total de segmentos y para *porcentajeCortesSegmentos*, se contabiliza el número de intersecciones totales entre el total de segmentos.

Además, el atributo *númeroRamas*, con relevancia nula estudiando las características de forma univariada, obtiene un resultado aceptable, pero sólo según *Random Forest*. Asimismo, el resto de atributos detectados anteriormente sin relevancia también son incluidos ahora como relevantes. *relaciónEjes* mejora, pero sólo con el método *ReliefF*. *mediaÁreaÁreaConvexa* empeora en gran medida, especialmente según *ReliefF*.

Los parámetros *excentricidad* y *relaciónEjes*, los cuales tenían la misma valoración entre ellos según Chi-cuadrado y Relación de Ganancia de Información, pasan a tener una valoración parecida aunque diferente según *ReliefF* y *Random Forest*.

Entre las diferencias destacables dentro de los dos métodos multivariables, *feretMínimoMáximo* pasa de la séptima posición según *ReliefF*, a la posición 42 según *Random Forest*. Por último, *númeroPartículasAgregado* obtiene una valoración destacable en todos los métodos menos en *ReliefF*.

2.6.2 Conjunto TEM

Para el tratamiento de las imágenes del conjunto TEM se ha seguido del mismo modo el algoritmo descrito en el apartado 2.3. En este caso, la configuración se ha variado ligeramente. En la fase de *binarización* tampoco se ha aplicado ningún suavizado preliminar. El filtro de suavizado de difusión anisotrópica [PM90] proporcionaba una segmentación más precisa en un número elevado de imágenes, sin embargo no se ha utilizado ya que se descartaban algunos

Tabla 2.4: Relevancia de los atributos según los métodos a) Chi-cuadrado y b) Relación de Ganancia de Información sobre el conjunto de datos TEM.

(a) Chi-cuadrado

Ranking	Atributo
567,088	absorción
567,088	númeroPartículasAgregado
567,088	circularidad
550,815	factorDeAgregación
522,801	perímetroConvexoPerímetro
516,742	relaciónÁreaÁreaConvexa
490,281	porcentajeSegmentos
489,207	tamañoPartícula
488,386	porcentajeCortesSegmentos
449,385	áreaÁreaBoundingBox
397,059	cambioÁreaSuavizada
326,277	perímetro
273,410	factorOclusión
269,211	ejeMayorElipse
260,894	feretMáximo
260,384	áreaConvexa
255,532	feret90
251,398	feretMínimo
248,368	mediaÁreaÁreaConvexa
243,486	perímetroFractal
242,563	desviaciónÁreaÁreaConvexa
240,309	ejeMenorElipse
219,387	diámetroEquivalente
219,387	volumenEsfera
219,387	área
196,020	excentricidad
196,020	relaciónEjes
162,712	volumenAgregado
154,756	feretMínimoMáximo
148,534	relaciónFeret90FeretMáximo
148,459	mediaÁreaÁreaTriángulo
142,943	desviaciónÁreaÁreaTriángulo
83,232	númeroRamas
67,537	perímetroÁrea
60,106	diámetroPartícula
60,106	volumenPartícula
42,886	desviaciónRamas
42,354	sectoresVacíos
39,402	centroideX
36,541	cambiosDerivada
36,125	mediaCentralRamas
34,475	mediaRecortadaRamas
33,466	medianaRamas
26,940	mediaRamas

(b) Relación Ganancia Información

Ranking	Atributo
0,316	factorDeAgregación
0,311	absorción
0,311	númeroPartículasAgregado
0,311	circularidad
0,299	relaciónÁreaÁreaConvexa
0,237	porcentajeCortesSegmentos
0,228	tamañoPartícula
0,227	áreaÁreaBoundingBox
0,210	cambioÁreaSuavizada
0,209	porcentajeSegmentos
0,205	perímetroConvexoPerímetro
0,191	mediaÁreaÁreaConvexa
0,178	perímetro
0,175	feret90
0,168	feretMínimo
0,165	feretMínimoMáximo
0,160	ejeMenorElipse
0,150	factorOclusión
0,142	feretMáximo
0,142	desviaciónÁreaÁreaConvexa
0,141	áreaConvexa
0,138	ejeMayorElipse
0,135	desviaciónRamas
0,130	relaciónEjes
0,130	excentricidad
0,126	perímetroFractal
0,121	volumenEsfera
0,121	diámetroEquivalente
0,121	área
0,121	volumenAgregado
0,120	desviaciónÁreaÁreaTriángulo
0,112	mediaÁreaÁreaTriángulo
0,110	diámetroPartícula
0,110	volumenPartícula
0,100	relaciónFeret90FeretMáximo
0,066	perímetroÁrea
0,058	centroideX
0,056	cambiosDerivada
0,047	númeroRamas
0,046	sectoresVacíos
0,036	mediaRecortadaRamas
0,035	mediaCentralRamas
0,033	medianaRamas
0,027	mediaRamas

agregados. El umbral de *binarización* ha sido determinado mediante el método Otsu [Ots75]. En la fase de filtrado de ruido, se han eliminado los elementos que no alcanzaban los 100 píxeles de área. Además, el disco empleado para el dilatado y la erosión ha sido de 8 píxeles de radio. El EMSA (Área Superficial

del Microscopio Electrónico) obtenido ha sido de 27,16 m^2/g, la mitad que con el conjunto SEM.

En la Tabla 2.4 se evalúa la relevancia de los atributos mediante los métodos Chi-cuadrado y Relación de Ganancia de Información sobre el conjunto de datos TEM. Como era de esperar, al tratarse de los mismos 45 atributos, hay grandes similitudes con los resultados de los mismos métodos para el conjunto de datos SEM. Aun así, su relevancia es diferente porque cada conjunto de datos contiene agregados diferentes.

En los 9 atributos mejor valorados de los conjuntos SEM y TEM, se encuentran las mismas características pero en diferente orden, a excepción de *mediaÁreaÁreaConvexa*, que para las imágenes SEM se encuentra en 4ª y 2ª posición según los métodos Chi-cuadrado y Relación de Ganancia de Información respectivamente. En cambio, para el conjunto TEM, ésta se encuentra en las posiciones 19ª y 12ª. En su lugar, dentro de los 10 mejores, en el conjunto TEM se encuentra *factorDeAgregación* en 3ª posición mientras que en el conjunto SEM se encuentra en la 11ª. El resto de los 9 atributos se encuentran ordenados de manera similar, con cierta variación para *porcentajeSegmentos* y *perímetroConvexoPerímetro*.

Cabe destacar también que en este caso, el único atributo con relevancia nula es $centroideY$. Esta reducción tan drástica de atributos nulos se debe a dos motivos. Por una parte, al ser mayor el conjunto de datos, de 266 agregados del conjunto SEM, a 781 en el TEM, es más fácil para los algoritmos encontrar relaciones entre los atributos y la clase morfológica. Por otra parte, la naturaleza del conjunto de datos también influye en la valoración de las mismas características. Si reducimos el conjunto TEM al mismo tamaño que el conjunto SEM, obtenemos 10 atributos con relevancia nula frente a los 13 generados mediante el conjunto SEM para los métodos Chi-cuadrado y Relación de Ganancia de Información. Así, puede observarse la vinculación entre el tamaño del conjunto de datos y la capacidad de relacionar las características con la categoría morfológica.

Destaca la mejora del atributo *cambioÁreaSuavizada*, aportado por esta investigación. Este parámetro realiza un proceso de suavizado y analiza el cambio en la relación ($área/áreaConvexa$) antes y después del suavizado. Además de las diferencias morfológicas de los agregados, las características de cada técnica de microscopía electrónica y el ajuste de los parámetros de binarización de ellas hacen susceptible la utilidad de este parámetro dependiendo de lo rugoso que quede el perímetro del agregado.

Tabla 2.5: Relevancia de los atributos según los métodos a) *ReliefF* y b) *Random Forest* sobre el conjunto de datos TEM.

(a) *ReliefF*

Ranking	Atributo
0,080	circularidad
0,078	factorDeAgregación
0,076	relaciónÁreaÁreaConvexa
0,073	númeroPartículasAgregado
0,062	porcentajeSegmentos
0,061	áreaÁreaBoundingBox
0,055	tamañoPartícula
0,048	perímetroConvexoPerímetro
0,046	perímetroFractal
0,046	porcentajeCortesSegmentos
0,043	perímetro
0,042	relaciónEjes
0,041	feretMáximo
0,041	feret90
0,039	diámetroPartícula
0,038	feretMínimo
0,038	ejeMenorElipse
0,037	feretMínimoMáximo
0,037	sectoresVacíos
0,037	absorción
0,037	diámetroEquivalente
0,036	ejeMayorElipse
0,035	cambioÁreaSuavizada
0,034	relaciónFeret90FeretMáximo
0,034	áreaConvexa
0,033	desviaciónÁreaÁreaTriángulo
0,032	mediaÁreaÁreaConvexa
0,029	excentricidad
0,029	desviaciónÁreaÁreaConvexa
0,028	área
0,021	factorOclusión
0,020	volumenPartícula
0,018	mediaÁreaÁreaTriángulo
0,018	volumenEsfera
0,017	centroideX
0,015	volumenAgregado
0,013	perímetroÁrea
0,013	mediaRecortadaRamas
0,010	númeroRamas
0,008	cambiosDerivada
0,008	desviaciónRamas
0,003	centroideY
0,001	mediaCentralRamas
0,001	medianaRamas
0,001	mediaRamas

(b) *Random Forest*

Ranking	Atributo
70,551	porcentajeSegmentos
68,118	porcentajeCortesSegmentos
66,581	factorDeAgregación
63,764	circularidad
63,764	absorción
63,636	númeroPartículasAgregado
62,868	perímetroConvexoPerímetro
61,716	relaciónÁreaÁreaConvexa
61,716	tamañoPartícula
58,259	áreaÁreaBoundingBox
57,746	cambioÁreaSuavizada
56,082	factorOclusión
53,649	perímetro
53,393	áreaConvexa
53,009	feretMáximo
52,753	númeroRamas
52,497	feretMínimo
52,369	feret90
52,241	ejeMayorElipse
51,857	ejeMenorElipse
51,472	mediaÁreaÁreaConvexa
50,832	sectoresVacíos
50,320	desviaciónÁreaÁreaConvexa
49,808	volumenAgregado
49,424	mediaÁreaÁreaTriángulo
49,168	diámetroEquivalente
49,040	volumenEsfera
49,040	área
48,784	perímetroFractal
46,223	relaciónEjes
46,223	excentricidad
45,455	desviaciónÁreaÁreaTriángulo
43,790	feretMínimoMáximo
43,406	relaciónFeret90FeretMáximo
42,894	centroideX
42,382	volumenPartícula
42,254	diámetroPartícula
41,613	mediaRamas
41,613	desviaciónRamas
41,229	mediaCentralRamas
40,461	perímetroÁrea
39,693	medianaRamas
39,053	mediaRecortadaRamas
37,772	cambiosDerivada
35,595	centroideY

En cuanto al resto de parámetros aportados en esta investigación, *mediaÁreaÁreaConvexa* ya hemos mencionado que ha empeorado notoriamente. *desviaciónÁreaÁreaConvexa* tiene una valoración parecida entre los métodos

Chi-cuadrado y Relación de Ganancia de Información. A diferencia de lo que ocurría en el conjunto SEM, *sectoresVacíos* mantiene la pobre valoración que tenía y *cambiosDerivada* deja de tener una valoración nula. Las características *porcentajeSegmentos* y *porcentajeCortesSegmentos* empeoran un poco pero siguen manteniendo una posición destacada.

Los atributos *excentricidad* y *relaciónEjes* vuelven a tener el mismo valor pero sufren una drástica reducción.

Por otra parte, en la Tabla 2.5 se muestran los resultados obtenidos para el conjunto TEM según los métodos *ReliefF* y *Random Forest*. El atributo *porcentajeSegmentos* sigue estando en cabeza, pero se ven ciertas diferencias respecto a los mismos métodos para el conjunto SEM. Las características *relaciónÁreaÁreaConvexa* y *áreaÁreaBoundingBox* pierden importancia. Todavía en mayor medida lo hacen *excentricidad* y *relaciónEjes*.

Con respecto a los métodos univariables, mantienen grandes similitudes en los 11 mejores atributos, a excepción de *absorción* y *cambioÁreaSuavizada* que empeoran para los métodos multivariables.

El atributo *númeroRamas* mejora notoriamente con el algoritmo *Random Forest*, incluso algo más que con el conjunto SEM.

Entre los dos métodos univariables, se dan grandes diferencias en los parámetros *absorción, cambioÁreaSuavizada, áreaConvexa* y *factorOclusión* que mejoran para *Random Forest* y en los parámetros *relaciónEjes, diámetroPartícula* y *feretMínimoMáximo* que obtienen mejor valoración según *ReliefF*.

Para el conjunto TEM, *excentricidad* y *relaciónEjes* reciben una valoración diferente según *ReliefF*, al igual que los dos métodos multivariables con SEM, pero no es así según *Random Forest* con el conjunto TEM.

2.6.3 Conjunto artificial

Para el tratamiento de las imágenes del conjunto artificial se ha seguido del mismo modo el algoritmo descrito en el apartado 2.3. En este caso, las imágenes, al haberse creado de manera artificial, están exentas de ruido. Por lo tanto, en la fase de *binarización* no se ha aplicado ningún suavizado preliminar. El umbral de *binarización* ha sido determinado mediante el método Otsu [Ots75]. En la fase de filtrado de ruido, no ha sido necesario eliminar los elementos que no alcanzaban un área determinada. Además, tampoco se ha llevado a cabo ni el dilatado ni la erosión. El EMSA (Área Superficial del Microscopio Electrónico) obtenido ha sido de 37,43 m^2/g, un valor intermedio

Tabla 2.6: Relevancia de los atributos según los métodos a) Chi-cuadrado y b) Relación de Ganancia de Información sobre el conjunto de datos artificial.

(a) Chi-cuadrado

Ranking	Atributo
1848,265	absorción
1848,265	circularidad
1848,265	númeroPartículasAgregado
1810,367	factorDeAgregación
1579,960	relaciónÁreaÁreaConvexa
1507,370	tamañoPartícula
1491,253	relaciónEjes
1491,253	excentricidad
1436,900	mediaÁreaÁreaConvexa
1380,720	feretMínimoMáximo
1297,950	desviaciónÁreaÁreaConvexa
1271,989	perímetroConvexoPerímetro
1257,743	relaciónFeret90FeretMáximo
1205,592	porcentajeCortesSegmentos
1126,160	áreaÁreaBoundingBox
1046,740	porcentajeSegmentos
981,177	cambioÁreaSuavizada
948,328	ejeMayorElipse
929,271	diámetroPartícula
929,271	volumenPartícula
871,273	perímetroFractal
857,897	feretMáximo
839,944	mediaÁreaÁreaTriángulo
728,780	perímetro
672,311	desviaciónRamas
644,084	factorOclusión
638,448	sectoresVacíos
448,455	númeroRamas
428,022	medianaRamas
400,572	mediaRamas
392,885	áreaConvexa
385,123	mediaCentralRamas
348,071	feret90
322,770	mediaRecortadaRamas
307,267	ejeMenorElipse
286,983	centroideX
279,902	feretMínimo
277,809	centroideY
263,521	desviaciónÁreaÁreaTriángulo
220,626	perímetroÁrea
189,822	volumenEsfera
189,822	diámetroEquivalente
189,822	área
116,583	volumenAgregado
50,770	cambiosDerivada

(b) Relación Ganancia Información

Ranking	Atributo
0,309	relaciónÁreaÁreaConvexa
0,308	factorDeAgregación
0,291	circularidad
0,291	absorción
0,291	númeroPartículasAgregado
0,264	áreaÁreaBoundingBox
0,254	porcentajeCortesSegmentos
0,245	mediaÁreaÁreaConvexa
0,241	tamañoPartícula
0,237	excentricidad
0,237	relaciónEjes
0,219	feretMínimoMáximo
0,216	porcentajeSegmentos
0,214	perímetroConvexoPerímetro
0,207	sectoresVacíos
0,201	relaciónFeret90FeretMáximo
0,200	desviaciónÁreaÁreaConvexa
0,186	diámetroPartícula
0,186	volumenPartícula
0,184	ejeMayorElipse
0,179	cambioÁreaSuavizada
0,178	mediaÁreaÁreaTriángulo
0,171	perímetroFractal
0,162	centroideX
0,145	factorOclusión
0,140	feretMáximo
0,137	perímetro
0,110	númeroRamas
0,110	desviaciónRamas
0,097	áreaConvexa
0,095	mediaRamas
0,082	ejeMenorElipse
0,082	feretMínimo
0,081	medianaRamas
0,076	desviaciónÁreaÁreaTriángulo
0,072	perímetroÁrea
0,071	área
0,071	diámetroEquivalente
0,071	volumenEsfera
0,069	mediaRecortadaRamas
0,069	feret90
0,067	centroideY
0,065	mediaCentralRamas
0,062	cambiosDerivada
0,048	volumenAgregado

con respecto a los conjuntos SEM y TEM. Para estimar esta valor hemos establecido un tamaño de píxel de 8,13 $nm/píxel$ a partir de un agregado similar a otros reales con un diámetro de feret de 650 nm.

Para el conjunto de datos artificial, según los cuatro métodos de selección de atributos, ninguna de las características ha obtenido una relevancia nula. Como se había explicado con el conjunto TEM, el tamaño del conjunto de datos, así como la naturaleza de los conjuntos de datos influye en la diferente valoración de los agregados que en el conjunto SEM habían resultado nulos. De hecho, la valoración de los atributos relacionados con la esqueletización mejora en el conjunto artificial.

En la Tabla 2.6 se presentan los resultados obtenidos con el conjunto artificial mediante los métodos Chi-cuadrado y Relación de Ganancia de Información. En cuanto al método Chi-cuadrado, se encuentran grandes similitudes con los resultados obtenidos con el conjunto TEM en los atributos mejor calificados: *absorción, circularidad, númeroPartículasAgregado, factorDeAgregación, relaciónÁreaÁreaConvexa* y *tamañoPartícula*. A su vez, hay otros atributos que obtienen mayores similitudes respecto al conjunto de datos SEM, como *relaciónEjes, excentricidad, perímetroConvexoPerímetro* y *mediaÁreaÁreaConvexa*.

En cuanto al resto de atributos presentados en esta investigación, *porcentajeSegmentos* y *porcentajeCortesSegmentos* empeoran respecto a los otros conjuntos de datos. De los dos atributos obtiene una mejor valoración general *porcentajeCortesSegmentos*. Éste, contabiliza todos los cortes de los segmentos aleatorios, mientras que *porcentajeSegmentos* sólo tiene en cuenta si cada segmento intersecciona con el agregado o no. En SEM sucede al revés y en TEM no existe el mismo acuerdo entre los diferentes algoritmos.

Al igual que con los conjuntos de datos SEM y TEM, entre los métodos de selección de atributos Chi-cuadrado y Relación de Ganancia de Información, los atributos mejor valorados coinciden y son más notorias las diferencias a medida que se disminuye su ranking. Aunque estas diferencias son menos notorias en el conjunto de datos artificial, las mayores disparidades se dan en las características *sectoresVacíos*, $mediaWindsorizadaRamas$ y $centroideX$.

Los métodos multivariables *ReliefF* y *Random Forest* se muestran en la Tabla 2.7. Según *ReliefF* el método más relevante es *factorDeAgregación* y a continuación se encuentra el atributo *relaciónEjes*, con un puesto lejano a los obtenidos con el resto de conjuntos, incluso con el mismo método. De modo contrario, *absorción* y *númeroPartículasAgregado* obtienen una valoración pésima respecto al resto de métodos de selección de atributos. En concreto, *absorción* se encuentra en la posición 33 según *ReliefF*, cuando éste es el mejor según los métodos Chi-cuadrado y *Random Forest*. Con el atributo *númeroPartículasAgregado* ocurre algo similar, bajando de la posición 27 con

Tabla 2.7: Relevancia de los atributos según los métodos a) *ReliefF* y b) *Random Forest* sobre el conjunto de datos artificial.

(a) *ReliefF*

Ranking	Atributo
0,099	factorDeAgregación
0,089	relaciónEjes
0,079	circularidad
0,074	feretMínimoMáximo
0,071	relaciónFeret90FeretMáximo
0,069	sectoresVacíos
0,066	desviaciónÁreaÁreaConvexa
0,066	excentricidad
0,057	relaciónÁreaÁreaConvexa
0,052	mediaÁreaÁreaConvexa
0,050	diámetroPartícula
0,048	áreaÁreaBoundingBox
0,048	desviaciónÁreaÁreaTriángulo
0,047	perímetroConvexoPerímetro
0,046	porcentajeSegmentos
0,041	cambioÁreaSuavizada
0,040	perímetroFractal
0,039	tamañoPartícula
0,037	centroideX
0,036	mediaÁreaÁreaTriángulo
0,034	ejeMayorElipse
0,033	porcentajeCortesSegmentos
0,032	ejeMenorElipse
0,031	feret90
0,030	feretMáximo
0,028	perímetro
0,028	númeroPartículasAgregado
0,028	feretMínimo
0,028	diámetroEquivalente
0,025	númeroRamas
0,023	cambiosDerivada
0,020	volumenPartícula
0,020	absorción
0,019	perímetroÁrea
0,019	áreaConvexa
0,019	centroideY
0,018	área
0,017	mediaRamas
0,017	mediaCentralRamas
0,017	medianaRamas
0,016	volumenAgregado
0,011	volumenEsfera
0,011	desviaciónRamas
0,010	factorOclusión
0,002	mediaRecortadaRamas

(b) *Random Forest*

Ranking	Atributo
49,441	absorción
49,441	númeroPartículasAgregado
49,385	circularidad
48,098	feretMínimoMáximo
47,987	relaciónFeret90FeretMáximo
47,707	porcentajeCortesSegmentos
47,595	factorDeAgregación
47,483	relaciónÁreaÁreaConvexa
46,197	porcentajeSegmentos
45,470	feretMáximo
44,183	excentricidad
44,127	relaciónEjes
43,344	tamañoPartícula
43,065	mediaÁreaÁreaConvexa
41,890	desviaciónÁreaÁreaConvexa
41,275	perímetroConvexoPerímetro
40,380	áreaÁreaBoundingBox
40,324	sectoresVacíos
39,429	cambioÁreaSuavizada
38,870	volumenPartícula
38,814	diámetroPartícula
38,647	mediaÁreaÁreaTriángulo
37,248	númeroRamas
36,242	perímetroFractal
36,130	ejeMayorElipse
35,514	perímetro
34,340	feretMínimo
32,942	medianaRamas
32,942	factorOclusión
32,494	feret90
31,823	áreaConvexa
31,656	desviaciónRamas
30,313	mediaRamas
29,866	mediaCentralRamas
29,251	diámetroEquivalente
29,251	área
28,971	volumenEsfera
28,915	centroideX
28,915	desviaciónÁreaÁreaTriángulo
28,803	ejeMenorElipse
28,635	centroideY
27,741	mediaRecortadaRamas
27,461	volumenAgregado
26,174	perímetroÁrea
25,280	cambiosDerivada

el método *ReliefF* a las primeras con el resto de métodos. Los atributos *circularidad*, *feretMínimoMáximo* y *relaciónFeret90FeretMáximo* se encuentran en las posiciones 3, 4 y 5 respectivamente en ambos métodos.

Resumiendo, el método *ReliefF*, presenta unos resultados diferentes al resto de técnicas de selección de atributos. Las características *circularidad* y *factorDeAgregación* obtienen buenos resultados según los cuatro métodos.

2.6.4 Tiempos

A continuación, en la Tabla 2.8 se muestra el tiempo requerido por cada método de selección de atributos para cada uno de los tres conjuntos de datos. Nótese que para el método de selección de atributos basado en el clasificador *Random Forest*, se muestran los tiempos para 10, 100 y 1.000 árboles. Este estudio de tiempos se ha llevado a cabo porque con el método que más se adapta al paso posterior de clasificación, el *Random Forest* de 1.000 árboles, a medida que el conjunto de datos aumenta de tamaño, el tiempo requerido para evaluar los atributos se incrementa exponencialmente. Por esto, se muestran los tiempos para los casos de 10 y 100 árboles para que se tengan en cuenta en los supuestos en los que el tiempo sea un problema o los conjuntos de datos sean mucho más grandes.

Tabla 2.8: Tiempo requerido en segundos en evaluar los atributos con cada uno de los tres conjuntos de datos. Junto con el nombre se indica el número de muestras. Para *Random Forest* se muestran los tiempos para 10, 100 y 1.000 árboles.

Conjunto de datos	$\boldsymbol{\chi^2}$	**Relación G.I.**	***ReliefF***	***Random Forest***		
				10	100	1.000
SEM (200)	0,737	0,808	0,900	8,244	60,285	624,869
TEM (781)	1,356	0,943	2,317	21,695	194,573	2093,411
Artif. (1.788)	1,564	1,137	8,230	55,974	549,973	5937,895

Los tiempos para los métodos χ^2 (Chi-cuadrado), Relación de Ganancia de Información y *ReliefF* son extremadamente bajos, sin embargo la valoración del método embebido *Random Forest*, se adapta mejor a la tarea de clasificación [SIL07].

2.7 Discusión de los resultados

Los resultados expuestos demuestran que las características que hemos diseñado en esta investigación son relevantes para diferenciar a los agregados y

categorizarlos. Además, se presenta la relevancia de los atributos empleados en la literatura, algo no publicado con anterioridad a esta investigación. A pesar de esto, los métodos de selección de atributos se han utilizado con éxito en multitud de campos [GE03, KPK+12, FKHC+12].

Uno de los objetivos de esta investigación es evaluar lo relacionado que está cada atributo con la clase, lo que justifica la elección de los dos métodos univariados: Chi-cuadrado y Relación de Ganancia de Información, frecuentemente usados en la literatura [SIL07]. Entre los métodos univariados se dan similitudes, así como entre los multivariados debido al desconocimiento o aprovechamiento de las interacciones entre los atributos, respectivamente.

Según los métodos univariados, el mejor atributo es *relaciónÁreaÁreaConvexa*, característica comúnmente usada [PI97, Rus07]. Fijándonos en el método Chi-cuadrado, se encuentra a continuación el atributo *áreaÁreaBoundingBox*, utilizado en la literatura para el tratamiento de imagen pero no en concreto para el *carbon black*.

En cambio, los multivariables son más completos por tener en cuenta las interacciones entre variables, que a fin de cuentas van a ser aprovechadas por los clasificadores. Siendo el objetivo principal de la tarea de selección de atributos el mejorar la clasificación, es necesario prestarles especial atención.

Por un lado, las características *porcentajeSegmentos* y *porcentajeCortesSegmentos* obtienen unos resultados increíbles. En concreto, para el conjunto TEM según el *Random Forest*, son los dos mejores parámetros, superando a todos los existentes en la literatura. En general, *mediaÁreaÁreaConvexa* y *cambioÁreaSuavizada* obtienen un resultado satisfactorio en la mayoría de los casos. Posteriormente, con un resultado todavía aceptable se encuentran los parámetros *desviaciónÁreaÁreaConvexa, mediaÁreaÁreaTriángulo* y *desviaciónÁreaÁreaTriángulo*. Por último, 6 de los parámetros propios han obtenido una relevancia casi nula. Éstos han sido *sectoresVacíos, cambiosDerivada* y los relacionados con las ramas: *desviaciónRamas, medianaRamas, mediaRecortadaRamas* y *mediaWindsorizadaRamas*.

La baja valoración que reciben los parámetros relacionados con las ramas indican que es necesario optimizar este proceso, ya que ha sido utilizado con éxito en la literatura [MG99] y, es claro, que de este tipo de características se puede extraer información morfológica. Se han realizado diversos ajustes para optimizar el proceso de esqueletización, pero queda trabajo por realizar en este ámbito.

Centrándonos en atributos concretos, hay varios que requieren especial

mención o recordatorio. Así, la característica *relaciónEjes* obtiene el mismo valor que *excentricidad* para los conjuntos SEM, TEM y Artificial con los métodos univariables Chi-cuadrado y Relación de Ganancia de Información. En cambio, al analizar su relevancia teniendo en cuenta las interacciones con el resto de atributos se desvela que su relevancia es diferente.

En general, los parámetros *circularidad, factorDeAgregación* y *número-PartículasAgregado* obtienen resultados muy buenos, a excepción del método *ReliefF* para los conjuntos SEM y artificial. El atributo *absorción* obtiene también una baja valoración con *ReliefF*, especialmente con el conjunto de datos artificial, en el que se encuentra en la posición 33. Sin embargo, éste es el mejor según los métodos Chi-cuadrado y *Random Forest* para el conjunto artificial y obtiene buenos resultados con los otros dos conjuntos y los otros tres métodos de selección de atributos.

El conjunto artificial, creado para ampliar el número de conjuntos de datos y poder perfeccionar y validar la herramienta, ha resultado ser más parecido al conjunto TEM que al SEM. En el capítulo 3, sobre la categorización de los agregados, se verá si este conjunto puede ser utilizado para ampliar el conjunto de datos de entrenamiento y mejorar la clasificación.

En cuanto a los tiempos, son mucho más elevados los requeridos por *Random Forest*, y en concreto el de 1.000 árboles, sin embargo siguen siendo más que aceptables, especialmente para una tarea que normalmente no requiere demasiada urgencia, como puede ser la reducción de la dimensionalidad de un conjunto para la creación de un modelo de clasificación. Sin embargo, es posible que en ciertos casos se prefiera reducir ligeramente la calidad a cambio de disminuir notablemente el tiempo, como pude ser con grandes conjuntos de datos o en caso de querer encadenar el proceso de selección de atributos con la generación de un modelo de clasificación frecuentemente actualizado.

Para concluir, cabe recalcar que los métodos de selección de atributos son una herramienta útil para evaluar su relevancia y ayudar en las tareas de diseño de nuevas características. De esta forma se han validado los parámetros diseñados en la presente investigación, al igual que se ha hecho en otros campos [KPK+12, FKHC+12].

2.8 Sumario

En este capítulo se han realizado tres importantes contribuciones. Se ha creado un método robusto de extracción de características, se han diseñado nuevos

atributos y se han introducido métodos de evaluación de características por primera vez en esta área.

Así, este capítulo ha presentado el estado de la técnica en la fabricación y caracterización del *carbon black*, así como en el tratamiento de imagen y los métodos de selección de atributos.

Además, se ha explicado el proceso realizado para la extracción de características de los agregados de *carbon black* para el cual se han optimizado los parámetros de procesamiento para las técnicas SEM y TEM. Adicionalmente, se ha creado un conjunto de datos artificial con el objetivo de evitar el problema de los datos no balanceados [EHG07].

En concreto, el proceso de segmentación ha sido diseñado de forma más robusta, para detectar el mayor número de agregados en diferentes condiciones, frente a buscar la segmentación óptima de agregados concretos, consiguiendo elaborar dos buenos conjuntos de datos a partir de imágenes con alta cantidad de ruido. Así, se ha llevado a cabo un análisis exhaustivo de diferentes filtros para eliminar ruido, como el suavizado gaussiano, el filtro de difusión anisotrópica y el filtro Wiener, y diferentes configuraciones del algoritmo de procesamiento de imágenes diseñado. A parte de su estudio individual, con las combinaciones de estos filtros y configuraciones se han procesado múltiples imágenes. Esto se debe a lo influenciados que están, por ejemplo, un filtro para la eliminación de ruido afecta al tamaño de disco óptimo para el dilatado y erosión de la imagen.

Asimismo, se han presentado los resultados de diferentes algoritmos de selección de atributos para evaluar la calidad de éstos, y se han comparado los existentes en la literatura con los diseñados en el presente trabajo. Se han utilizado dos métodos univariables y dos multivariables y se han estudiado las implicaciones de ambos métodos. Así, a la hora de analizar un parámetro durante su proceso de diseño puede ser útil la valoración de un algoritmo univariable, pero si el objetivo final es la inclusión en un algoritmo de aprendizaje automático, es necesario tener en cuenta las interacciones entre variables, de las que este algoritmo se va a aprovechar.

En cuanto a los parámetros diseñados en esta investigación, destacan *porcentajeSegmentos* y *porcentajeCortesSegmentos*, que superan según el método de evaluación de la relevancia de los atributos con el algoritmo *Random Forest* y el conjunto TEM al resto de los atributos, incluso a los establecidos por la literatura.

Resumiendo, se ha presentado un método para evaluar las características

de los nanomateriales y se ha probado que las aportadas en esta investigación han resultado relevantes.

«A los hombres se les puede dividir en dos categorías: los que hablan para decir algo, y los que dicen algo por hablar.»

Príncipe Carlos José de Ligne
(1735-1814)

3

Categorización del *carbon black*

Actualmente, para extraer información de los agregados de *carbon black* se han extendido los métodos indirectos, ya descritos en el apartado 2.1.5.1. Aun así, en esta investigación hemos optado por los métodos directos, ya que éstos son capaces de proporcionar la distribución de las categorías morfológicas.

Los métodos directos extraen información de los agregados de un nanomaterial capturado por un microscopio. A partir de esta información se han desarrollado diferentes métodos de clasificación en categorías [HMH92, MG99]. Estos enfoques, mediante análisis discriminante, obtienen ecuaciones para realizar esta categorización. Nosotros hemos decidido aprovechar el potencial de los algoritmos de aprendizaje automático [LdRS$^+$10], no usado con anterioridad en este ámbito y además, en el capítulo 2 hemos ampliado el número de características descritas en la literatura.

En concreto, las contribuciones que hemos realizado en este ámbito son las siguientes:

- Proporcionamos y evaluamos un método de clasificación de agregados de *carbon black* mediante algoritmos de aprendizaje automático.

- Discutimos la problemática de no tener las clases balanceadas y aportamos diferentes soluciones.

El resto del capítulo queda distribuido de la siguiente forma. Se comienza con una breve introducción de la problemática existente en la sección 3.1. Después, la sección 3.2 detalla los 5 algoritmos de aprendizaje automático empleados. Posteriormente, en la sección 3.3, se explica la importancia del análisis ROC para evaluar la bondad de un clasificador. A continuación, en la sección 3.4, se presentan los experimentos de categorización en 4 clases morfológicas sobre tres conjuntos de datos. Seguidamente, en la sección 3.5, se discuten los resultados obtenidos, y por último, en la sección 3.6, se resumen las contribuciones del capítulo.

3.1 Introducción

Llegados a este punto, ha quedado clara la relevancia de la morfología de los agregados de *carbon black*. Es por esto que en esta tesis se ha elegido mejorar los métodos existentes de análisis directo de imágenes de microscopio. En esta investigación, se ha confiado en su potencial y se han querido mejorar a pesar de haber sido desplazados por los métodos indirectos, como la absorción de aceite y la adsorción de nitrógeno [WN09].

Cabe recalcar que los estudios existentes [MH72, HMSH93, MG99] no proporcionan valores como *accuracy* ni AUC para evaluar la calidad de los métodos utilizados para clasificar los agregados. Únicamente, indican que las clases son significativamente diferentes mediante el método T2 de Hotelling [HMSH93].

Asimismo, los métodos descritos en la literatura no son exactamente replicables. Por esto, en este apartado se evalúan diferentes algoritmos de aprendizaje automático sobre diferentes conjuntos de datos, pero no son comparados con los métodos existentes.

El estado de la técnica presentado en este capítulo pertenece al concepto de la Inteligencia Artificial. Según John McCarthy, "es la ciencia e ingeniería de hacer máquinas inteligentes" [McC07]. Comenzó a usarse en los años 50, y hoy en día tiene multitud de ramas, entre las que se encuentra el aprendizaje automático, expuesto en el apartado 3.2. Una de las técnicas con más importancia para la evaluación de los métodos de aprendizaje automático es el análisis de la curva ROC [Met78], descrito en el apartado 3.3. Además, otra

de las ramas principales de la IA son los algoritmos genéticos, la base del algoritmo de reconstrucción 3D propuesto en el capítulo 4.

3.2 Aprendizaje Automático

Las características obtenidas a través del procesado 2D y 3D de las imágenes SEM y TEM de las nano-partículas, así como el resultado de inspecciones de estos nanomateriales, son una fuente de información que puede ser utilizada para conseguir predecir las características mecánicas de los materiales. De este modo, el aprendizaje automático o *machine learning* es una rama de la inteligencia artificial que trata de modelar el espacio formado por datos de entrenamiento, de tal modo que ante la aparición de una nueva instancia sea posible clasificarlo entre las diferentes clases existentes [Bis06].

En general, los algoritmos de aprendizaje automático pueden clasificarse en tres tipos: supervisados, no-supervisados y semi-supervisados. En primer lugar, los algoritmos supervisados, requieren que el conjunto de datos de entrenamiento esté debidamente etiquetado; esto es, que cada instancia perteneciente a ese conjunto de datos haya sido correctamente clasificada con anterioridad [Bis06]. En segundo lugar, los algoritmos no-supervisados tratan de resolver cómo están organizados los datos de entrenamiento. Se distingue del aprendizaje supervisado en que los datos no se encuentran etiquetados [KP04]. Por último, los algoritmos de aprendizaje semi-supervisado utilizan una mezcla entre datos etiquetados y datos no etiquetados, mejorando así la precisión de los modelos no supervisados [CSZ06].

Teniendo en cuenta las necesidades del modelo propuesto, así como la capacidad subyacente de etiquetar apropiadamente los datos, consideramos que el enfoque correcto para la clasificación y análisis de los datos, es el del aprendizaje supervisado. Por lo tanto, en lo que resta de capítulo revisaremos algunos de los enfoques más comunes dentro del aprendizaje supervisado, que han tenido éxito en casos similares [NSP+09, SNPB09]. Así, en el apartado 3.2.1, se explica el funcionamiento de las redes bayesianas. Posteriormente, en el apartado 3.2.2, se expone la base de los SVM, en el apartado 3.2.3, se explica el método de los vecinos más próximos (KNN) y finalmente, en el apartado 3.2.4 se expone el funcionamiento de los árboles de decisión.

3.2.1 Redes Bayesianas

Las redes bayesianas son modelos de aprendizaje automático supervisado, basados en el teorema de Bayes [Bay63]. De acuerdo con su formulación clásica, mostrada en la ecuación (3.1), dados dos eventos A y B, la probabilidad condicional $P(A|B)$, es decir, de que A ocurra si B ha ocurrido puede obtenerse si sabemos la probabilidad de que A ocurra $P(A)$, la probabilidad de que B ocurra $P(B)$ y la probabilidad condicional de B dado A, $P(B|A)$.

$$P(A \mid B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.1}$$

Extendiendo este enfoque, las redes bayesianas son modelos probabilísticos para el análisis multivariable. Se pueden representar como un grafo acíclico dirigido y la distribución de probabilidad asociada a ese grafo [CGH97]. Por un lado, el modelo gráfico representa las relaciones probabilísticas entre las diferentes variables que representan un problema. Por otro lado, la función de probabilidad muestra la fuerza de las relaciones o arcos en el grafo.

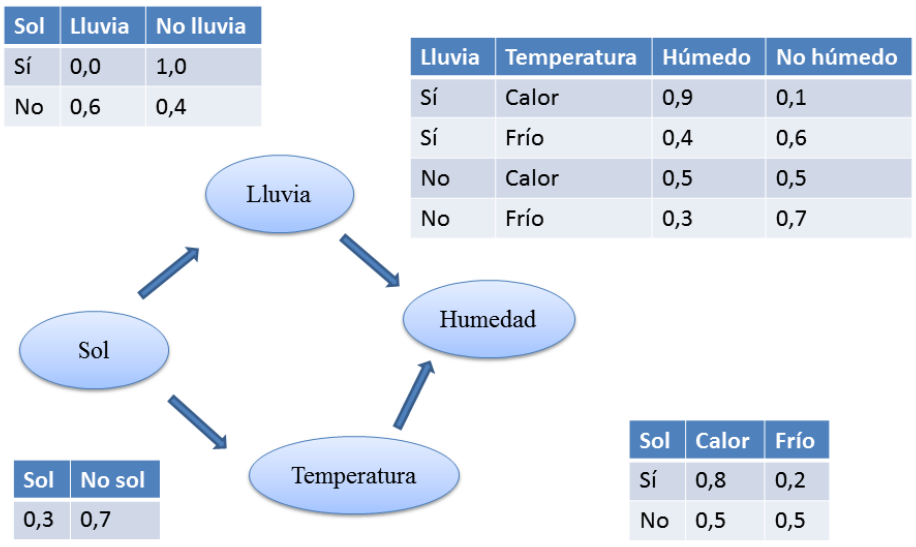


Figura 3.1: Ejemplo de una red bayesiana.

La Figura 3.1 muestra un ejemplo de red bayesiana. Las flechas entre los nodos indican la inter-dependencia entre variables, que es la probabilidad condicionada. Las tablas que se incluyen en la imagen son las distribuciones de probabilidad condicionada.

De este modo, utilizando este tipo de modelos podemos no sólo inferir un resultado con una cierta significancia (por ejemplo, si va a llover ó no), sino también identificar las causas de un determinado problema, actuando las redes bayesianas como una caja transparente a aquel que las utilice. Además, las redes bayesianas son capaces de dar un resultado, aun cuando algunos de los valores de las variables de entrada estén incompletos [CGH97].

Este modelo se puede utilizar con distintos fines, como el aprendizaje automático sobre datos históricos, reconocimiento de patrones sobre datos incompletos o ambiguos, minería de datos para reconocer relaciones e inferencia de variables no observables, dado el resto de datos [PR00]. En concreto, la capacidad de inferencia representa un conjunto semántico de los denominados sistemas expertos que están basados en el encadenamiento de reglas tanto hacia delante como hacia atrás (en realidad, los sistemas bayesianos permiten una tercera forma de inferencia, que se conoce como explicación o justificación [CGH97]). Además, una red bayesiana puede crecer ampliando su conocimiento base con nuevas evidencias sin reducir su precisión [CGH97] mientras se adapta al problema y mantiene un procedimiento actualizado.

3.2.2 *Support Vector Machines* (SVM)

Los SVM, o *máquinas de vectores de soporte* son modelos de aprendizaje automático que se utilizan para clasificación y regresión [Bis06]. A partir de los datos de entrenamiento los SVM construyen un modelo para que cuando haya que clasificar un nuevo ejemplo, se sepa a qué clase pertenece. Así, los SVM construyen un *hiperplano* o conjunto de *hiperplanos* en un espacio n-dimensional. En la Figura 3.2 se muestra un SVM bidimensional. Además, la separación conseguida por el hiperplano es la distancia más grande a los puntos de datos más cercanos (llamada esta distancia margen funcional) [Vap99].

Formalmente, partiendo de un conjunto de datos de entrenamiento que es de la forma $\mathcal{D} = \{\mathbf{x_i}, c_i \mid \mathbf{x_i} \in \mathbb{R}^\rho, c_i \in \{-1, 1\}\}_{i=1}^n$ donde c_i es -1 ó 1, siendo este valor el indicativo de la clase a la que el vector $\mathbf{x_i}$ pertenece. Cada $\mathbf{x_i}$ es un vector ρ-dimensional formado por números reales. Lo que se quiere encontrar es el hiperplano con mayor margen que divida a los puntos con $c_i = 1$ de los $c_i = -1$. Se puede encontrar cualquier hiperplano compuesto de puntos $\mathbf{x}$ siempre que se satisfaga que $\mathbf{w} \cdot \mathbf{x} - b = 0$, donde $\cdot$ denota el producto escalar entre ambos vectores, el vector $\mathbf{w}$ que es el vector normal, esto es, perpendicular al hiperplano y el parámetro $\frac{b}{\|\mathbf{w}\|}$ indica el desplazamiento del hiperplano respecto al origen a través del vector normal.

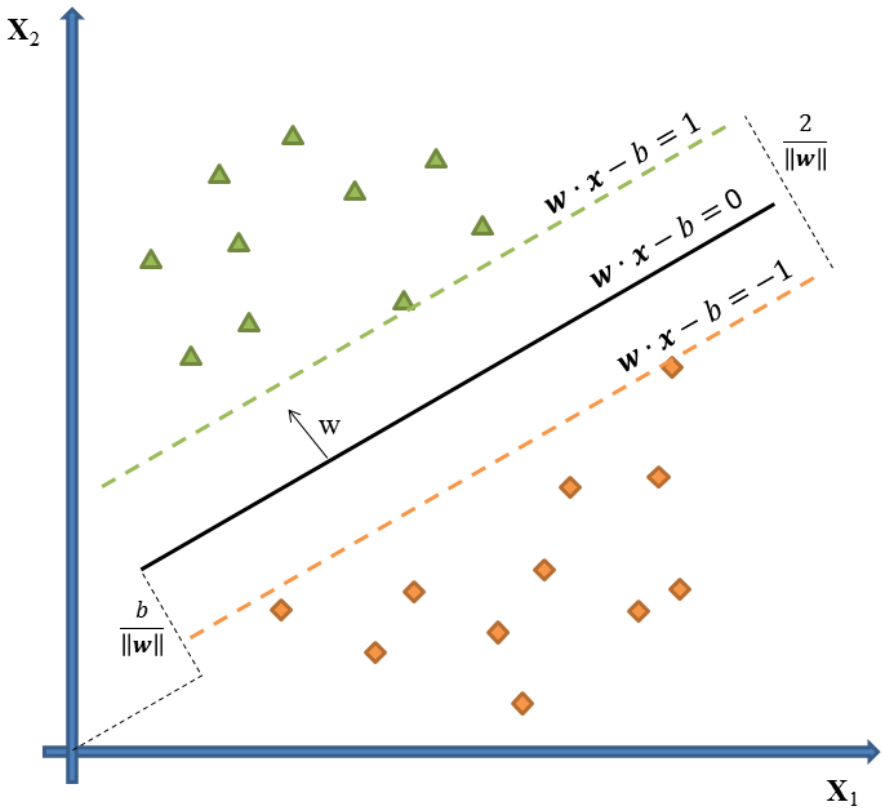


Figura 3.2: Ejemplo de un SVM bidimensional.

Se quiere escoger el vector $\mathbf{w}$ y la b que maximicen la distancia entre los hiperplanos paralelos, es decir, queremos hallar el margen óptimo. De este modo, estos dos hiperplanos serían, por un lado, $\mathbf{w} \cdot \mathbf{x} - b = 1$ para la clase $c_i = 1$ y, por otro lado, $\mathbf{w} \cdot \mathbf{x} - b = -1$ para la clase $c_i = -1$. Si el conjunto de datos de entrenamiento es linealmente separable, se pueden seleccionar dos hiperplanos que cumplan estas condiciones. Utilizando geometría básica, encontramos que la distancia entre los dos hiperplanos es $\frac{2}{\|\mathbf{w}\|}$, por lo tanto, se quiere minimizar $\|\mathbf{w}\|$. Además, como también se quiere prevenir que no existan puntos entre estos dos hiperplanos, se añade la siguiente restricción: para cada i, $\mathbf{w} \cdot \mathbf{x_i} - b \geq 1$ para $\mathbf{x_i}$ de la primera clase y $\mathbf{w} \cdot \mathbf{x_i} - b \leq -1$ para $\mathbf{x_i}$ de la segunda. Esto se puede rescribir como $c_i(\mathbf{w} \cdot x_i - b) \geq 1$ para todo i de la forma $1 \leq i \leq n$, llegando así al problema de optimización del margen: optimizar en $\mathbf{w}, b, \|\mathbf{w}\|$ sujeto a que para todo i de la forma $1 \leq i \leq n$, se cumpla $c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1$.

Resolviendo el problema utilizando métodos cuadráticos se llega a que el

desplazamiento se define como:

$$b = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} (\mathbf{w} \cdot \mathbf{x_i} - c_i) \tag{3.2}$$

donde N_{sv} es el número de vectores que tiene un factor de multiplicación resultante de la resolución cuadrática α_i, mayor que 0, y se llaman *vectores de soporte*, y el vector normal $\mathbf{w}$ se define como:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i c_i \mathbf{x_i} \tag{3.3}$$

Concretamente, el modelo de hiperplano que se ha presentado es un SVM de *kernel* lineal. Sin embargo, normalmente se utilizan *kernels* más complejos para dividir el espacio de entrenamiento.

Muchas veces por el simple hecho de que el espacio de entrenamiento no es linealmente separable. El algoritmo es similar, sólo que se modifican los productos escalares por funciones de *kernel*. Este cambio consigue que la división resultante se ajuste mejor en el espacio de características resultante.

3.2.3 *K-nearest-neighbours* (KNN)

El algoritmo KNN [FHJ52] es uno de los algoritmos más sencillos de aprendizaje automático supervisado. Este algoritmo está basado en la clase de las instancias más cercanas de una instancia a clasificar.

Concretamente, en la fase de entrenamiento de este clasificador se representa un conjunto de datos de entrenamiento de la forma $\mathcal{D} = \{\mathbf{x_i}, c_i \mid \mathbf{x_i} \in \mathbb{R}^n\}$, dónde $\mathbf{x_i}$ es un vector n-dimensional y c_i es la clase de ese vector. De este modo, en la fase de clasificación de una instancia desconocida $\mathbf{x_m}$ se lleva a cabo midiendo la distancia de ese $\mathbf{x_m}$ respecto a los datos ya almacenados en el modelo. Para esto, se suele utilizar la distancia euclídea, que se define como:

$$d = \sqrt{\sum_{i=1}^{n} (\mathbf{x_i} - \mathbf{x_m})^2} \tag{3.4}$$

Así, en la Figura 3.3, se muestra un espacio bidimensional en el que se encuentran dos tipos de instancias: triángulos y rombos. El parámetro K es

el número de instancias vecinas a tener en cuenta para clasificar una instancia desconocida, en este caso, el pentágono. Empleando 5 vecinos el pentágono es clasificado como triángulo, ya que hay tres triángulos frente a dos rombos. En cambio, si nos fijamos en los 11 vecinos más próximos, hay 6 rombos frente a 5 triángulos, y el pentágono sería clasificado como rombo.

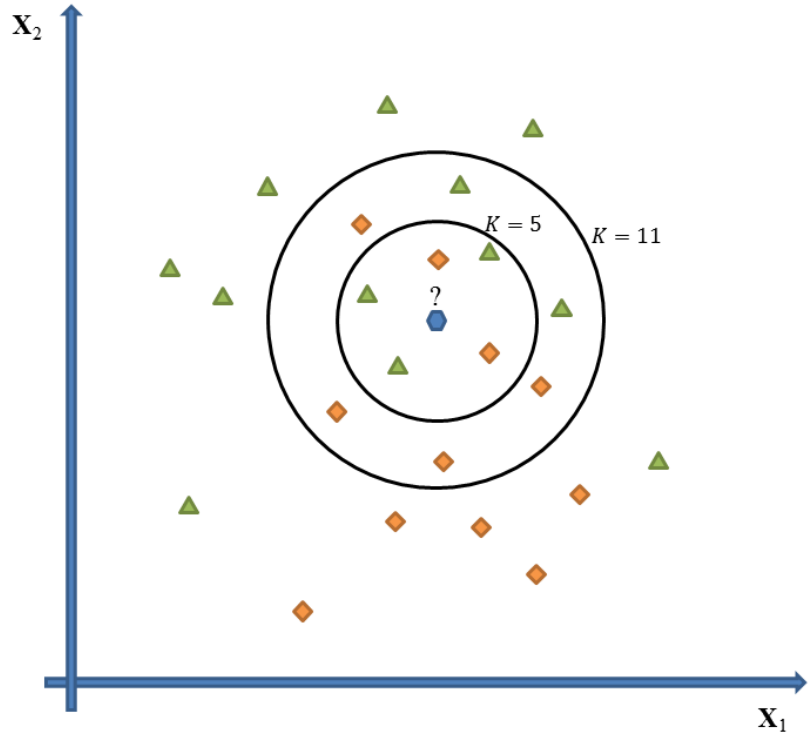


Figura 3.3: Ejemplo de un clasificador KNN.

3.2.4 Árboles de decisión

Los árboles de decisión o *Decision Trees* son una herramienta de soporte a la decisión que usa un gráfico o modelo de las decisiones en forma de árbol [Qui86b]. Este modelo incluye las decisiones posibles, así como las posibles consecuencias, las probabilidades de que un evento se dé, el coste de los recursos y su utilidad. De esta forma se genera un modelo de *caja blanca*, ya que dado el resultado es posible obtener una simple explicación matemática del mismo [Bis06].

Formalmente, el gráfico del árbol de decisión $G = (V, E)$ consiste en un conjunto V de nodos finitos y no-vacíos y un conjunto de aristas E.

Si el conjunto de aristas son pares ordenados (v, w) de vértices, entonces el gráfico es dirigido. Un camino es una secuencia de aristas en la forma $(v_1, v_2), (v_2,\ v_3),\ \ldots,\ (v_{n-1},\ v_n)$. Se dice que el camino es de v_1 *a* v_n, y su distancia es de n. Si (v, w) es una arista del árbol, v es considerado como el padre y w como el hijo de v. Aquel nodo que no tiene padres es denominado *nodo raíz*. Por el contrario, aquellos nodos que no tienen descendencia son considerados como *nodos terminales*. Todos los demás son denominados *nodos internos*.

La construcción de un árbol de decisión se basa en varios elementos: un conjunto de preguntas binarias Q de la forma $\{x\ \in A\}$ donde A es un subconjunto del espacio muestral, un método para separar los nodos, una estrategia para detener el crecimiento del árbol y la asignación de cada nodo terminal a un valor de la variable de respuesta, para realizar una regresión o una clase para la clasificación. Con todo esto, las diferencias entre los algoritmos se encuentran en la poda de los árboles, en la regla para separar los nodos y en el tratamiento de los valores perdidos.

En la Figura 3.4 se muestra un árbol de decisión con una característica numérica, la *edad*, y dos características binarias, ser *fumador* y la calidad de la *dieta*. Las dos clases posibles son *riesgo* y *no riesgo* de padecer cáncer. El funcionamiento de este modelo de ejemplo consiste en empezar por el nodo superior y comprobar el valor de la característica *fumador*. Si es afirmativo pasamos al nodo izquierdo y comprobamos el valor de la *edad*. Si es menor de 30 años, quedaría clasificado como un paciente *sin riesgo*, en cambio, si es mayor de 30 años, tendría *riesgo* de padecer cáncer. En el caso de que el individuo no sea *fumador*, la *dieta* determinaría el *riesgo* de padecer cáncer. En el caso de ser buena, no tendría *riesgo*, y en el caso de ser mala sí.

Existen multitud de algoritmos para implementar los árboles de decisiones. En 1984, Breiman *et ál.* introdujeron una algoritmo de árboles de decisión binario denominado CART [BFO$^+$84], utilizando como criterio de partición la impureza del nodo. Por otro lado, Kass en 1980 introdujo CHAID (*Chi-square automatic interaction detection*) [Kas80] como derivado del THAID (*A sequential analysis program for the analysis of nominal scale dependent variables*) de Morgan y Messenger [MM73], utilizando como criterio para separar basado en χ^2 y para terminar el proceso se debe definir de antemano un umbral. Los algoritmos ID3 [Qui79, Qui83] y C4.5 [Qui93] formulados por Quinlan proveen de un algoritmo simple pero potente que es capaz de trabajar con valores continuos y discretos.

Los árboles de decisión se suelen utilizar en operaciones de análisis de la

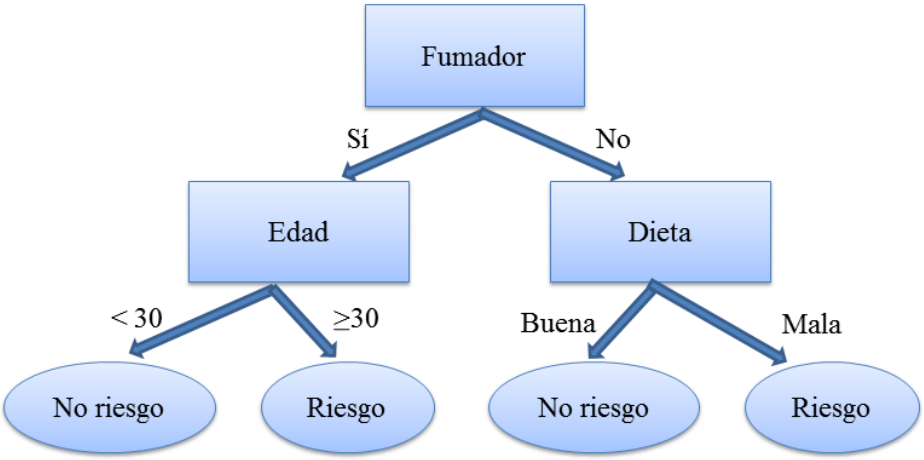


Figura 3.4: Ejemplo de un árbol de decisión.

decisión [PCRR+04, AMG12], para ayudar a identificar una estrategia con más posibilidades de alcanzar una meta. Una variante de los árboles de decisión es el bosque aleatorio o *Random Forest* [Bre01]. Éste es un clasificador formado por muchos árboles de decisión de tal forma que la clasificación final es la más votada por los árboles individuales. Estos árboles se caracterizan por tener la misma distribución de clases.

3.3 Análisis ROC

El porcentaje de aciertos es la medida que a primera vista puede parecer suficiente para evaluar un clasificador, sin embargo, esto no es así y conviene que esté acompañada de otras medidas, como el área por debajo de la curva ROC (AUC) [Met78]. Esto es especialmente cierto en conjuntos de datos no balanceados [Faw06], es decir, con un número de elementos diferente de cada clase. Así, ante un cambio en la proporción de las instancias de cada clase, la curva ROC no se verá afectada [Faw06], a diferencia de otras métricas como el porcentaje de aciertos. Un ejemplo para comprenderlo es el caso en el que una enfermedad rara se da en el 2 % de los pacientes. Un clasificador que clasifique siempre al individuo como sano tendrá un porcentaje de aciertos del 98 % dando una falsa apariencia de su calidad.

Para comprender el análisis mediante la curva ROC (Característica Operativa del Receptor, en inglés *Receiver Operating Characteristic*) es necesario

conocer los conceptos de sensibilidad y especificidad, que a su vez requieren del conocimiento de los verdaderos y falsos positivos y de los verdaderos y falsos negativos.

Este tipo de análisis va dirigido a clasificaciones binarias, como enfermo / sano, bueno / malo, en cambio el problema que se trata en esta tesis es el de una clasificación en 4 categorías morfológicas. La solución consiste en crear una curva ROC por cada clase [Faw06]. A continuación, para explicar los diferentes términos se utiliza la clase ramificada, como clase positiva, frente al resto, como clase negativa.

Al clasificar un elemento hay cuatro resultados posibles [NM10]:

1. Verdadero positivo (VP). El elemento es ramificado y el clasificador lo etiqueta correctamente.

2. Falso negativo (FN). El elemento es ramificado, pero el clasificador lo etiqueta como de otra clase.

3. Falso positivo (FP). El elemento no es ramificado y el clasificador lo etiqueta incorrectamente como ramificado.

4. Verdadero negativo (VN). El elemento no es ramificado y el clasificador lo etiqueta correctamente como de otra clase.

El total de positivos (P) está formado por los verdaderos positivos y los falsos negativos. Del mismo modo, el total de negativos (N) está formado por los falsos positivos y los verdaderos negativos. En la Tabla 3.1 se muestra la matriz de confusión o tabla de contingencia, una herramienta de gran utilidad para identificar como son los errores de un clasificador.

		Clase real	
		p	**n**
Resultado predicción	**Sí**	VP	FP
	No	FN	VN
	Total:	**P**	**N**

Tabla 3.1: Matriz de confusión.

Aprovechando estos conceptos, la sensibilidad (véase la ecuación (3.5)) queda determinada como la probabilidad de que un elemento ramificado se

clasifique correctamente, lo que es decir, es la capacidad del clasificador para que, ante un agregado ramificado, sea capaz de determinar su clase correctamente. Se calcula dividiendo el número de agregados correctamente clasificados como ramificados entre el total de agregados ramificados existentes en el conjunto de datos. También es denominada como tasa de aciertos, Fracción de Verdaderos Positivos (FVP), en inglés *True Positive Rate* o *recall*.

$$Sensibilidad = \frac{VP}{VP + FN} = \frac{VP}{P} = FVP \tag{3.5}$$

Del mismo modo, la especificidad o fracción de verdaderos negativos (véase la ecuación (3.6)) queda determinada como la probabilidad de que un elemento no ramificado se clasifique correctamente, lo que es decir, es la capacidad del clasificador para que, ante un agregado no ramificado, sea capaz de determinarlo como tal. Se estima dividiendo el número de agregados correctamente clasificados como no ramificados entre el total de agregados no ramificados existentes en el conjunto de datos.

$$Especificidad = \frac{VN}{VN + FP} = \frac{VN}{N} = 1 - FFP \tag{3.6}$$

donde la FFP (Fracción de Falsos Positivos) queda determinada por la relación entre los falsos positivos, es decir, los negativos mal clasificados, y el total de negativos, como puede verse en la ecuación (3.7). La FFP también se conoce como tasa de falsa alarma:

$$FFP = \frac{FP}{N} \tag{3.7}$$

Por otro lado, existen otros dos conceptos, valor predictivo positivo y valor predictivo negativo, con una orientación más práctica, ya que se plantean ante el desconocimiento del valor real. Así, el valor predictivo positivo (VPP) (véase la ecuación (3.8)) refleja la probabilidad de que un agregado clasificado como ramificado realmente lo sea, es decir, representa la fiabilidad del algoritmo ante un agregado que ha sido clasificado como ramificado. Se calcula dividiendo el número de agregados correctamente clasificados como ramificados entre el total de agregados clasificados como ramificados.

$$VPP = \frac{VP}{VP + FP} \tag{3.8}$$

Del mismo modo, el valor predictivo negativo (VPN) (véase la ecuación (3.9)) refleja la probabilidad de que un agregado clasificado como no ramificado realmente no sea ramificado, es decir, representa la fiabilidad del algoritmo ante un agregado que ha sido clasificado como no ramificado. Se estima dividiendo el número de agregados correctamente clasificados como no ramificados entre el total de agregados clasificados como no ramificados.

$$VPN = \frac{VN}{FN + VN} \tag{3.9}$$

Otros dos términos de gran importancia y que es importante diferenciar son la precisión y la exactitud. La precisión (en inglés *precision*) o valor predictivo positivo (véase la ecuación (3.10)) se estima dividiendo el número de agregados correctamente clasificados como ramificados entre el total de agregados clasificados como ramificados.

$$Precisión = \frac{VP}{VP + FP} \tag{3.10}$$

La exactitud (en inglés *accuracy*) o porcentaje de aciertos (véase la ecuación (3.11), se estima dividiendo el total de agregados correctamente clasificados entre el total de agregados existentes en el conjunto de datos:

$$Exactitud = \frac{VP + VN}{P + N} \tag{3.11}$$

Área por debajo de la curva ROC (AUC)

La curva ROC fue concebida durante la segunda guerra mundial para determinar si un punto en la pantalla del radar representaba a un objeto enemigo o a ruido [WFU06]. Es una forma de visualizar gráficamente el rendimiento de un clasificador. El área por debajo de la curva ROC (AUC) es la probabilidad de que una instancia positiva elegida aleatoriamente sea correctamente clasificada con mayor sospecha que una negativa también elegida aleatoriamente [ZHH$^+$09].

Como puede verse en la Figura 3.5 se representa como la sensibilidad frente a (1 – especificidad), o lo que es lo mismo, FVP frente a FFP, para un sistema clasificador binario según se varía el umbral de discriminación. Proporciona valores entre 0 y 1, siendo 0.5 el valor obtenido con un clasificador aleatorio.

Existen dos tipos de clasificadores, los discretos, que solo proporcionan un valor de clasificación, como los árboles de decisión y los probabilísticos, que

Inst#	Clase	Puntuación	Inst#	Clase	Puntuación
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

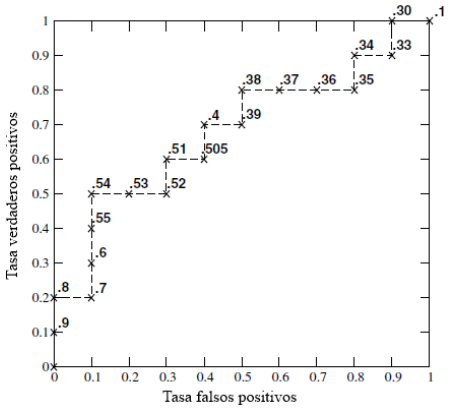


Figura 3.5: Curva ROC creada variando el umbral de las puntuaciones indicadas en la tabla. Basado en [WFU06].

proporcionan la probabilidad de que una instancia pertenezca a una clase, como una red neuronal. Con este segundo tipo de clasificadores se varía el umbral de dicha probabilidad para crear la curva ROC [Faw06]. En cambio, los clasificadores discretos sólo se representan mediante un punto en el espacio ROC para un conjunto de datos. Sin embargo, aunque no disponen de una probabilidad directa al clasificar una instancia, se puede utilizar información interna para generar un índice de confianza. Por ejemplo, los árboles de decisión determinan la clase de una hoja por la proporción de instancias en dicho nodo. Así, esta proporción será utiliza como la probabilidad de que dicha clasificación sea correcta.

Para el caso de una clasificación multiclase, una forma de calcular el AUC total es la media ponderada de los AUC de cada clase [PD00], como se indica en la ecuación (3.12):

$$AUC_{total} = \sum_{c_i \in C} AUC(c_i) \cdot p(c_i) \tag{3.12}$$

donde $p(c_i)$ es la prevalencia de la clase c_i, es decir, la proporción de elementos que pertenecen a dicha clase con respecto al total de elementos del conjunto de datos. En este caso, la opción elegida en este trabajo, el AUC deja de ser totalmente independiente de la distribución de clases [Faw06].

Otro enfoque para adaptar las curvas ROC a problemas multiclase independiente de la distribución de clases, mide la discriminabilidad de pares de clases no balanceados [HT01] (véase la ecuación (3.13)):

$$AUC_{total} = \frac{2}{|C|(|C|-1)} \sum_{\{c_i,c_j\} \in C} AUC(c_i, c_j) \tag{3.13}$$

donde $AUC(c_i, c_j)$ es el área por debajo de la curva ROC formada entre las clases c_i y c_j. El sumatorio se calcula para todos los pares de clases sin importar el orden, lo que resulta en un número de pares igual a $|C|(|C|-1)/2$. Este método tiene el problema de que no hay una forma fácil de visualizar la superficie del área que se está calculando.

3.4 Evaluación experimental

En esta sección se encuentran los experimentos llevados a cabo con el fin de validar la aplicación de los métodos de aprendizaje automático al área de la clasificación morfológica de los agregados de *carbon black*, algo no investigado previamente en la literatura.

3.4.1 Metodología general

Para la evaluación del método de categorización propuesto, inicialmente segmentamos los agregados de las imágenes, posteriormente los etiquetamos manualmente en 4 categorías morfológicas [HMH92]. A continuación, generamos un fichero ARFF (*Attribute-Relation File Format*) formato similar al CSV (*Comma Separated Values*) con todas las características descritas en el apartado 2.3.4 incluida la clasificación. Finalmente, realizamos estudios de aprendizaje automático (del inglés, *machine learning*) para clasificar los agregados mediante el *framework Weka* [Gar95, HFH+09, BFH+11].

Mediante esta experimentación se busca responder a las siguientes preguntas:

- *¿Qué algoritmo de aprendizaje proporciona mejores resultados?*
- *¿Se comportan todos los algoritmos de manera uniforme en los tres conjuntos de datos?*
- *¿Mejora la clasificación con un conjunto de datos balanceado, como es el conjunto de datos artificial?*

- *¿Mejora la clasificación de los conjuntos de datos reales ampliándolos con el conjunto de datos artificial?*
- *¿Cuál es el coste computacional de cada algoritmo?*

Para las tres primeras preguntas se empleará el *accuracy* o precisión de acierto y el AUC (Área por debajo de la curva ROC) y 17 clasificadores sobre los tres conjuntos de datos. Para la cuarta pregunta se ampliará cada uno de los conjuntos reales con el conjunto artificial para intentar mejorar la clasificación. Por último, para la quinta pregunta, se medirán los tiempos requeridos tanto para generar el modelo de aprendizaje automático, como para clasificar un agregado.

Específicamente, seguimos la siguiente metodología para probar la idoneidad de cada algoritmo de *machine learning*.

- ***Cross validation:*** este método se utiliza frecuentemente para evaluar la calidad de algoritmos de *machine learning* y evitar que los resultados se distorsionen por culpa del *overfitting* [Bis95]. En nuestros experimentos realizamos una validación cruzada con grupos de tamaño $k = 10$. De esta forma, el conjunto de datos se parte en 10 trozos de los que 9 son utilizados para el entrenamiento y 1 para la validación. El proceso se repite tantas veces como número de grupos hay, en este caso 10. Así, cada trozo acaba siendo utilizado una vez como validación.
- **Entrenando al modelo:** para cada grupo, realizamos el entrenamiento de cada modelo usando diferentes parámetros y algoritmos descritos en la sección 3.2). En particular usamos los siguientes modelos:
  - *Árboles de decisión (AD)*, del inglés *Decision Trees*: llevamos a cabo experimentos con el J48, la implementación de *Weka* [Gar95, HFH+09, BFH+11] del algoritmo *C4.5* [Qui93]) y *Random Forest* [Bre01], un conjunto de árboles de decisión construidos aleatoriamente con 10, 100 y 1.000 árboles.
  - *Support Vector Machines (SVM):* llevamos a cabo experimentos con un *kernel* polinomial [AW99], un *kernel* polinomial normalizado [MBM08] y un *kernel* universal basado en la función de *Pearson VII* [ÜMB07].
  - *K-nearest neighbour (KNN):* realizamos experimentos con número de vecinos k para $k = 1$, $k = 5$, $k = 10$, $k = 15$, $k = 20$ y $k = 25$ [FHJ52].

  - *Redes bayesianas*, del inglés *Bayesian Networks*: respecto a las redes bayesianas, hemos realizado experimentos con el clasificador *Naïve Bayes* [Lew98] así como con diferentes algoritmos estructurales de aprendizaje: *K2* [CH91], *Hill Climber* [RN03] y *Tree Augmented Naïve* (TAN) [GGP+97].

- **Validando los modelos:** evaluamos el porcentaje de instancias clasificadas correctamente y el área por debajo de la curva ROC (*Area Under the ROC Curve*, AUC) que establece la relación entre los falsos negativos y los falsos positivos [SKM09]. Comparamos el porcentaje de aciertos con el obtenido a partir de 6 clasificaciones manuales realizados por expertos en la materia. La tarea de categorización tiene un alto grado de subjetividad que la hemos medido comparando la clasificación individual con la clasificación por voto de los 6 expertos. Así, en 888 imágenes SEM y TEM, la media del porcentaje de acierto de las clasificaciones individuales es del 80 %. Con lo que si conseguimos realizar la misma tarea de manera automática con la misma efectividad se considerará un gran logro.

- **Análisis de tiempos:** medimos el tiempo requerido para generar el modelo del clasificador y validarlo, y por otro lado, el tiempo empleado en clasificar una instancia utilizando el modelo previamente guardado. A pesar de ser tareas que se realizan en espacios de tiempo reducidos, se ha considerado interesante presentar dicha información para ser tenida en cuenta en caso de utilizar grandes conjuntos de datos. Por la naturaleza de los ordenadores las mediciones de tiempo no son 100 % precisas, ya que hay tareas del sistema operativo que se ejecutan en segundo plano y no de manera uniforme. Para minimizar este impacto no deseado los modelos de los clasificadores se generan 20 veces para obtener la media del tiempo requerido. Además, para calcular el tiempo necesario para clasificar una instancia, se ha clasificado 20 veces el conjunto entero con el modelo guardado.

Para la evaluación de los tiempos de proceso se ha empleado un portátil ASUS® modelo G50V con procesador Intel® Mobile Core 2 Duo T5750 a 2 GHz y 4 GB de memoria RAM Doble-Canal DDR2 @ 332MHz. Todos los programas se han ejecutado bajo Windows® 7 Professional.

3.4.2 Conjunto SEM

En un primer estudio, para la conferencia *DEXA (International Conference on Database and Expert Systems Applications)* extrajimos 26 características de cada agregado del conjunto SEM [LdRS+10], descrito en el apartado 2.4.1. El conjunto de datos no estaba balanceado para las cuatro clases existentes y los datos eran escasos, 266 agregados. Específicamente, había 9 agregados de tipo esferoidal, 86 del tipo elipsoidal, 51 del lineal y 120 ramificados. Posteriormente, se realizó una purga adicional y un reetiquetado más sistemático para evitar la subjetividad de esta tarea. De esta forma, quedaron 4 de tipo esferoidal, 76 de tipo elipsoidal, 33 de tipo lineal y 87 ramificados. Gracias a esto, añadido a la eliminación de 9 atributos y la adición de 17 nuevos, se han mejorado los resultados de clasificación.

Tabla 3.2: Resultados de los algoritmos de aprendizaje automático con el conjunto de datos SEM.

Modelo de *Machine-learning*	***Accuracy* (%)**	**AUC**
AD: J48	75,50	0,810
AD: Random Forest con 10 árboles	74,50	0,909
AD: Random Forest con 100 árboles	**79,50**	0,926
AD: Random Forest con 1.000 árboles	78,50	**0,931**
SVM: Kernel Polinomial	71,50	0,828
SVM: Kernel Polinomial Normalizado	71,00	0,800
SVM: Kernel Universal Pearson VII	73,50	0,805
KNN K=1	66,50	0,737
KNN K=5	70,00	0,859
KNN K=10	71,00	0,892
KNN K=15	72,00	0,896
KNN K=20	72,50	0,908
KNN K=25	71,00	0,905
Bayes ingenuo	67,00	0,867
Red Bayesiana: K2	70,50	0,873
Red Bayesiana: Hill Climber	70,50	0,873
Red Bayesiana: TAN	72,50	0,890

En la Tabla 3.2 están registrados los resultados obtenidos en términos de porcentaje de aciertos (en inglés *accuracy*) y del área por debajo de la curva ROC (AUC), ya que, fijarse sólo en el porcentaje de aciertos puede ser equivocado, como se ha explicado en el apartado 3.3. En cuanto al *accuracy*, el

mejor algoritmo es el *Random Forest* de 100 árboles, con un 79,5 %, prácticamente el mismo porcentaje obtenido con la media de acierto de 6 expertos, que como hemos indicado anteriormente ha sido de un 80 %. En cambio, el *Random Forest* de 1.000 árboles es el que obtiene el mejor AUC, con un valor de 0,931. El peor algoritmo de aprendizaje automático es el KNN de 1 vecino, tanto en términos de porcentaje de aciertos, con un 66,5 %, como en cuanto el AUC, con un 0,737.

Para el mejor clasificador, considerando tanto el porcentaje de aciertos como el AUC, el algoritmo *Random Forest* de 1.000 árboles, en la Tabla 3.3 se muestra su matriz de confusión o tabla de contingencia para el conjunto de datos SEM. Esta herramienta, ya descrita en el apartado 3.3 sobre el análisis ROC, se utiliza para conocer con precisión los errores del clasificador. En la tabla podemos ver como el mayor número de errores se da al intentar clasificar los agregados lineales, ya que más de la mitad de éstos son clasificados incorrectamente. De los 33 agregados lineales sólo 12 son clasificados correctamente, 6 son clasificados como elipsoidales y 15 como ramificados. Así, queda presente la gran similitud morfológica existente entre ambas clases. De los 4 agregados esferoidales, sólo 1 ha sido clasificado correctamente y tres han sido clasificados como elipsoidales. A continuación, se encuentran los agregados elipsoidales, esta vez con sólo 13 % de fallos, de ellos, 2 agregados elipsoidales han sido confundidos con lineales y 8 como ramificados. Por último, los mejor clasificados han sido los ramificados, de los que también había más muestras para entrenar. De los 87, sólo 4 han sido confundidos con elipsoidales y 5 con lineales. De estos resultados, podemos concluir que es necesario diseñar algún atributo que sea capaz de diferenciar a los agregados lineales ya que es una tarea que no se consigue realizar satisfactoriamente.

Tabla 3.3: Matriz de confusión para el método *Random Forest* de 1.000 árboles aplicado sobre el conjunto de datos SEM.

a	b	c	d	← clasificado como
1	3	0	0	a = esferoidal
0	66	2	8	b = elipsoidal
0	6	12	15	c = lineal
0	4	5	78	d = ramificado

En cuanto al tiempo de procesamiento requerido, en la Tabla 3.4 se muestra para cada algoritmo, por un lado, (i) el tiempo necesario para generar el modelo y validarlo, y por otro, (ii) el tiempo empleado en clasificar una ins-

tancia utilizando el modelo previamente guardado. Como puede observarse, los tiempos para generar el modelo y validarlo son muy reducidos, a excepción del *Random Forest* de 100 árboles y especialmente largo el *Random Forest* de 1.000 árboles, que llega a tardar 20 veces más que el *Random Forest* de 10 árboles. Aun así, 38 segundos en generar el modelo es un tiempo más que aceptable para el clasificador que obtiene el mejor AUC y segundo mejor porcentaje de acierto, a poca diferencia del *Random Forest* de 100 árboles. Así, dependiendo del tamaño del conjunto de datos de entrenamiento que se utilice, habrá que elegir el algoritmo que se considere más apropiado. Además, también habrá que considerar cuál es el enfoque a seguir. Por un lado, se puede actualizar el modelo a medida que se tengan más datos, en cuyo caso, el tiempo de generación del modelo es bastante relevante. Por otro lado, se puede generar un modelo una sola vez, y utilizarlo para clasificar todas las nuevas instancias, en cuyo caso, el tiempo de generación del modelo no supondría un problema tan grande.

Tabla 3.4: Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos SEM.

Modelo de *Machine-learning*	**Generar modelo (s)**	**Clasificar instancia (s)**
AD: J48	2,32	$\mathbf{3{,}59{\times}10^{-3}}$
AD: Random Forest con 10 árboles	1,80	$5{,}44{\times}10^{-3}$
AD: Random Forest con 100 árboles	5,86	$1{,}23{\times}10^{-2}$
AD: Random Forest con 1.000 árboles	37,74	$5{,}06{\times}10^{-2}$
SVM: Kernel Polinomial	2,29	$6{,}15{\times}10^{-3}$
SVM: Kernel Polinomial Normalizado	3,84	$1{,}29{\times}10^{-2}$
SVM: Kernel Universal Pearson VII	5,45	$2{,}04{\times}10^{-2}$
KNN K=1	**0,92**	$4{,}05{\times}10^{-3}$
KNN K=5	1,01	$4{,}29{\times}10^{-3}$
KNN K=10	1,04	$4{,}25{\times}10^{-3}$
KNN K=15	1,08	$4{,}25{\times}10^{-3}$
KNN K=20	1,07	$4{,}36{\times}10^{-3}$
KNN K=25	1,12	$4{,}33{\times}10^{-3}$
Bayes ingenuo	1,33	$6{,}10{\times}10^{-3}$
Red Bayesiana: K2	1,49	$4{,}77{\times}10^{-3}$
Red Bayesiana: Hill Climber	1,92	$4{,}79{\times}10^{-3}$
Red Bayesiana: TAN	1,93	$5{,}43{\times}10^{-3}$

Volviendo a comparar el *Random Forest* de 10 árboles con el de 1.000, el se-

gundo tarda 9 veces más, aun así, para cada instancia sólo requiere $5{,}06 \times 10^{-2}$ segundos, o lo que es lo mismo, es capaz de clasificar 19 instancias por segundo. El clasificador más rápido en cuanto a generar el modelo y validarlo es el KNN de 1 vecino, sin embargo éste ha obtenido el peor porcentaje de acierto y AUC de todos los clasificadores. Este algoritmo tiene un tiempo de generación del modelo prácticamente nulo, ya que sólo consiste en almacenar las instancias. Además, hay que recordar que el tiempo registrado en la Tabla 3.4 incluye también el tiempo de validación. Por otro lado, el algoritmo más rápido en clasificar una instancia es el árbol de decisión J48 el cual obtenía unos resultados aceptables en términos de *accuracy* y AUC. El más lento vuelve a ser el *Random Forest* de 1.000 árboles.

3.4.3 Conjunto TEM

El segundo conjunto está formado por imágenes obtenidas a partir de un microscopio electrónico de transmisión, descrito en el apartado 2.4.2. El conjunto tampoco está balanceado y contiene 781 agregados, de los cuales 9 son esferoidales, 211 elipsoidales, 162 lineales y 399 ramificados.

En la Tabla 3.5 se muestran los resultados de los clasificadores en términos de porcentaje de aciertos y AUC sobre el conjunto de datos TEM. Destaca de nuevo el algoritmo *Random Forest* de 1.000 árboles, con incluso mejor valoración que con el conjunto SEM, con un 81,69 % de aciertos y un AUC de 0,94. Asimismo, supera la clasificación manual realizada con 80 % de precisión de acierto de media. La versión de 100 árboles y el *Support Vector Machine* con *kernel* polinomial le siguen de cerca, con la misma precisión de acierto, un 81,05 %, a pesar de que con el primer conjunto el SVM no obtenía resultados especialmente buenos. La red Bayesiana TAN y el KNN de 15 vecinos obtienen también una buena calificación. Según el AUC destacan, con más de un 0,9, los tres *Random Forest* y todos los KNN excepto el de 1 vecino. Recordemos, la ventaja del AUC respecto al porcentaje de aciertos para la evaluación de la calidad de un algoritmo, como se explicaba en el apartado 3.3 al describir la curva ROC, ésta no se ve afectada ante un cambio en la proporción de las instancias de cada clase, a diferencia de otras métricas como el porcentaje de aciertos [Faw06].

En cuanto a los peores clasificadores son de nuevo el Bayes ingenuo y el KNN de 1 vecino, tanto en términos de *accuracy* como de AUC. Las redes Bayesianas K2 y *Hill Climber* obtienen también un mal resultado, y los tres SVM, a pesar de obtener un buen porcentaje de aciertos no proporcionan un

Tabla 3.5: Resultados de los algoritmos de aprendizaje automático con el conjunto de datos TEM.

Modelo de *Machine-learning*	***Accuracy* (%)**	**AUC**
AD: J48	75,29	0,812
AD: Random Forest con 10 árboles	80,41	0,924
AD: Random Forest con 100 árboles	81,05	0,939
AD: Random Forest con 1.000 árboles	**81,69**	**0,940**
SVM: Kernel Polinomial	81,05	0,879
SVM: Kernel Polinomial Normalizado	80,92	0,870
SVM: Kernel Universal Pearson VII	80,54	0,868
KNN K=1	72,47	0,787
KNN K=5	77,85	0,903
KNN K=10	77,98	0,918
KNN K=15	78,11	0,919
KNN K=20	77,21	0,919
KNN K=25	77,08	0,920
Bayes ingenuo	71,83	0,892
Red Bayesiana: K2	73,37	0,895
Red Bayesiana: Hill Climber	73,37	0,895
Red Bayesiana: TAN	79,00	0,920

AUC aceptable.

Tabla 3.6: Matriz de confusión para el método *Random Forest* de 1.000 árboles aplicado sobre el conjunto de datos TEM.

a	b	c	d	← clasificado como
7	2	0	0	a = esferoidal
1	180	14	16	b = elipsoidal
0	22	89	51	c = lineal
0	20	17	362	d = ramificado

Sobre el conjunto TEM vuelve a ser el algoritmo *Random Forest* de 1.000 árboles el mejor valorado, esta vez tanto por el porcentaje de aciertos como por el AUC y en la Tabla 3.6 se muestra su matriz de confusión. Los agregados lineales vuelven a ser mal clasificados, siendo el 45 % de ellos mal clasificados, confundidos especialmente con los ramificados. Sin embargo, los agregados esferoidales pasan a la segunda posición, con sólo dos de ellos confundidos por

elipsoidales, pasan de un 75 % mal clasificados a un 22 %. A pesar de ser un gran cambio, su relevancia es limitada por la pequeña cantidad de agregados esferoidales, lo que lo convierte en poco representativo. Los elipsoidales, de los cuales el 15 % son clasificados incorrectamente, prácticamente son confundidos por igual entre lineales y ramificados, con 14 y 16 instancias respectivamente. Por último, los agregados ramificados son de nuevo los mejor clasificados, con sólo un 9 % de errores, así de los 399 agregados, 17 son confundidos con lineales y 20 con elipsoidales. En general, la clasificación ha mejorado a excepción de con los agregados elipsoidales con los que ha empeorado ligeramente, en concreto 2 puntos porcentuales.

Tabla 3.7: Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos TEM.

Modelo de *Machine-learning*	**Generar modelo (s)**	**Clasificar instancia (s)**
AD: J48	2,99	$\mathbf{1{,}43{\times}10^{-3}}$
AD: Random Forest con 10 árboles	3,22	$2{,}43{\times}10^{-3}$
AD: Random Forest con 100 árboles	19,17	$6{,}38{\times}10^{-3}$
AD: Random Forest con 1.000 árboles	165,70	$3{,}87{\times}10^{-2}$
SVM: Kernel Polinomial	2,74	$1{,}55{\times}10^{-3}$
SVM: Kernel Polinomial Normalizado	12,49	$7{,}20{\times}10^{-3}$
SVM: Kernel Universal Pearson VII	14,18	$9{,}43{\times}10^{-3}$
KNN K=1	2,03	$1{,}99{\times}10^{-3}$
KNN K=5	2,34	$2{,}24{\times}10^{-3}$
KNN K=10	2,50	$2{,}38{\times}10^{-3}$
KNN K=15	2,63	$2{,}48{\times}10^{-3}$
KNN K=20	2,69	$2{,}64{\times}10^{-3}$
KNN K=25	2,81	$2{,}71{\times}10^{-3}$
Bayes ingenuo	**1,75**	$1{,}74{\times}10^{-3}$
Red Bayesiana: K2	2,54	$1{,}97{\times}10^{-3}$
Red Bayesiana: Hill Climber	3,72	$2{,}05{\times}10^{-3}$
Red Bayesiana: TAN	3,75	$2{,}36{\times}10^{-3}$

En la Tabla 3.7 se muestra para cada algoritmo, por un lado, (i) el tiempo necesario para generar el modelo y validarlo, y por otro, (ii) el tiempo empleado en clasificar una instancia utilizando el modelo previamente guardado con el conjunto de datos TEM. Con respecto al conjunto de datos SEM, pasamos de 200 agregados a 781, lo que influye más en unos algoritmos de aprendizaje automático que en otros. El tiempo requerido por algunos algoritmos en

generar su modelo y validarlo se incrementa notoriamente en menor medida que el incremento de tamaño del conjunto de datos para los algoritmos: árbol de decisión J48, *Random Forest* de 10 árboles, *Support Vector Machine* con *kernel* polinomial, los 6 KNNs, Bayes ingenuo y las tres redes bayesianas. El resto de algoritmos suponen un coste de tiempo razonable, a excepción del *Random Forest* de 1.000 árboles, con el que se aprecia un crecimiento mayor que el aumento de instancias. El método más rápido es, en este caso, el Bayes ingenuo con 1,75 segundos, superando a todos los KNNs que eran los que habían obtenido los mejores resultados con el conjunto SEM, aun así, éstos obtienen de nuevo un buen tiempo. El mejor algoritmo en cuanto a porcentaje de acierto y AUC, el *Random Forest* de 1.000 árboles, obtiene otra vez el peor tiempo, separándose incluso más del resto de los algoritmos.

En cuanto al tiempo necesario para clasificar una sola instancia, el árbol de decisión J48 vuelve a ser el más rápido. En general los tiempos son inferiores a los requeridos con el conjunto SEM. Aunque a primera vista pueda parecer extraño, se debe a que el tiempo en clasificar una instancia es menor si los ficheros son más grandes, ya que se nota menos la sobrecarga que supone el leer un archivo. Para apreciar la forma en la que influye el tamaño del conjunto de datos, se presentan los tiempos requeridos en los tres conjuntos de datos. Además, en los algoritmos más rápidos, la diferencia es más notoria. El *Random Forest* de 1.000 árboles vuelve a ser el más lento y de los *Support Vector Machine* sólo destaca el de *kernel* polinomial.

3.4.4 Conjunto artificial

Este conjunto fue creado artificialmente para disponer de un conjunto adicional y balanceado del que poder extraer conclusiones y validar la herramienta. A partir de los resultados del KNN presentados en la Tabla 3.8 observamos diferencias con los conjuntos SEM y TEM. Los agregados artificiales no son tan diferentes entre sí, como los agregados originales, por lo que el KNN de 1 vecino obtiene en este caso un resultado muy bueno. El algoritmo KNN, K vecinos más próximos (del inglés, *K nearest neighbours*), consiste en buscar en el conjunto de entrenamiento, las K instancias más parecidas a la que se quiere clasificar y se decide su clase por mayoría. A partir del análisis de los resultados de los diferentes KNNs en los tres conjuntos de datos, podemos observar cómo con el conjunto artificial, al contrario de lo que sucede en los otros dos conjuntos, cuantos menos vecinos se tenga en cuenta mejor porcentaje de aciertos se obtiene. La razón de este comportamiento es la uniformidad de las

muestras que se da en el conjunto artificial, por lo cual con menos vecinos se consiguen mejores resultados.

Tabla 3.8: Resultados de los algoritmos de aprendizaje automático con el conjunto de datos artificial.

Modelo de *Machine-learning*	***Accuracy* (%)**	**AUC**
AD: J48	76,40	0,853
AD: Random Forest con 10 árboles	82,66	0,953
AD: Random Forest con 100 árboles	84,00	0,965
AD: Random Forest con 1.000 árboles	84,06	**0,966**
SVM: Kernel Polinomial	81,88	0,914
SVM: Kernel Polinomial Normalizado	80,15	0,908
SVM: Kernel Universal Pearson VII	**85,07**	0,931
KNN K=1	82,66	0,880
KNN K=5	81,66	0,951
KNN K=10	80,37	0,952
KNN K=15	79,53	0,950
KNN K=20	78,80	0,948
KNN K=25	78,19	0,947
Bayes ingenuo	69,24	0,911
Red Bayesiana: K2	75,67	0,927
Red Bayesiana: Hill Climber	75,67	0,927
Red Bayesiana: TAN	79,53	0,944

Gracias a esta mayor similitud entre agregados y estar las clases más balanceadas, se obtienen mejores resultados en todos los algoritmos en términos de AUC, y en cuanto al porcentaje de acierto, sólo es superado por los resultados del conjunto TEM con los algoritmos *Support Vector Machine* con *kernel* polinomial normalizado y con el Bayes ingenuo.

El mejor resultado con este conjunto se obtiene con el método *Support Vector Machine* con *kernel* universal *Pearson VII* según el porcentaje de acierto, con un 85,07 %, superando ampliamente la calidad de la clasificación media de los 6 expertos, que ha sido del 80 %. Le sigue el *Random Forest* de 1.000 árboles que obtiene prácticamente la misma valoración que el de 100 árboles. El de 10 árboles obtiene la misma calificación que el KNN de 1 vecino, algo totalmente diferente a lo que ocurría con los dos conjuntos anteriores, como se ha explicado al comienzo de este apartado. En concreto, tanto para SEM como para TEM la diferencia era de 8 puntos porcentuales.

En cambio, según el AUC vuelve a destacar el *Random Forest* de 1.000 árboles, seguido por los de 100 y 10 árboles y el KNN de 10 vecinos. El KNN de 1 vecino, a pesar de destacar por su porcentaje de acierto, no lo hace por su AUC. Por último, el árbol de decisión J48 es el que peor AUC obtiene.

Tabla 3.9: Matriz de confusión para el método *Support Vector Machine* con *kernel* universal *Pearson VII* aplicado sobre el conjunto de datos Artificial.

a	b	c	d	← clasificado como
379	17	1	2	a = esferoidal
43	353	47	20	b = elipsoidal
0	65	371	17	c = lineal
5	33	17	418	d = ramificado

En este caso, en la Tabla 3.9 se muestra la matriz de confusión para el algoritmo *Support Vector Machine* con *kernel* universal *Pearson VII*. Se pueden observar grandes diferencias respecto a las otras dos matrices presentadas. Para empezar, gracias a disponer de un número de agregados esferoidales equivalente a las otras categorías morfológicas, ésta es de hecho, la mejor clasificada, algo normal debido a la gran diferencia morfológica respecto a las otras tres clases. Sólo el 5 % de estos agregados son mal clasificados. Los agregados ramificados vuelven a ser bien categorizados con un 12 % de ellos mal clasificados, ligeramente superior a los dos conjuntos anteriores. En la tercera posición se encuentran los agregados lineales, con una buena clasificación en este conjunto. De los 82 fallos, 65 son de confundirlos con elipsoidales, a diferencia de los conjuntos SEM y TEM, en los que eran más confundidos con ramificados. Por último, los elipsoidales son los que resultan más difíciles de categorizar en este conjunto, con un 24 % de ellos mal clasificados y confundidos en una proporción similar entre esferoidales y lineales.

De nuevo se presentan los tiempos requeridos por cada algoritmo en la Tabla 3.10 para poder estudiar la forma en la que influye el tamaño del conjunto en los tiempos de procesamiento. En general los tiempos requeridos para generar los modelos aumentan proporcionalmente al número de muestras involucradas con respecto al conjunto TEM. El método *Random Forest* de 1.000 árboles tarda casi 8 minutos en generar el modelo y validarlo, pero tarda un tiempo similar al requerido por el conjunto TEM y algo inferior al SEM en clasificar una sola instancia.

Vuelve a destacar el Bayes ingenuo, tanto en generar el modelo como en clasificar cada instancia, con un valor increíblemente bajo, $8{,}45\times10^{-4}$. No

Tabla 3.10: Tiempos de los algoritmos de aprendizaje automático con el conjunto de datos artificial.

Modelo de *Machine-learning*	Generar modelo (s)	Clasificar instancia (s)
AD: J48	7,14	$8{,}80\times10^{-4}$
AD: Random Forest con 10 árboles	7,16	$1{,}57\times10^{-3}$
AD: Random Forest con 100 árboles	50,28	$4{,}99\times10^{-3}$
AD: Random Forest con 1.000 árboles	476,74	$3{,}43\times10^{-2}$
SVM: Kernel Polinomial	4,95	$8{,}85\times10^{-4}$
SVM: Kernel Polinomial Normalizado	43,68	$7{,}76\times10^{-3}$
SVM: Kernel Universal Pearson VII	41,07	$9{,}02\times10^{-3}$
KNN K=1	4,78	$1{,}83\times10^{-3}$
KNN K=5	6,44	$2{,}40\times10^{-3}$
KNN K=10	6,85	$2{,}55\times10^{-3}$
KNN K=15	7,65	$2{,}78\times10^{-3}$
KNN K=20	8,21	$2{,}88\times10^{-3}$
KNN K=25	8,33	$2{,}95\times10^{-3}$
Bayes ingenuo	**2,28**	$\mathbf{8{,}45\times10^{-4}}$
Red Bayesiana: K2	3,83	$1{,}07\times10^{-3}$
Red Bayesiana: Hill Climber	6,67	$1{,}23\times10^{-3}$
Red Bayesiana: TAN	6,66	$1{,}40\times10^{-3}$

obstante, este algoritmo arroja la peor clasificación en términos de porcentaje de aciertos. En cuanto al tiempo que tardan los KNNs en generar el modelo, se observa con el conjunto artificial un aumento mucho mayor al incrementar el número de vecinos.

3.5 Discusión de los resultados

Como hemos expuesto al describir la metodología de la evaluación experimental, nuestra meta era alcanzar el mismo porcentaje de acierto conseguido mediante una clasificación manual de 6 expertos, es decir, del 80 %. Consideramos el reto superado con un 79,5 % para el conjunto SEM, un 81,69 % para el conjunto TEM y un 85,07 % para el conjunto artificial.

Los resultados muestran que las características de los conjuntos, así como su tamaño, son determinantes para las tareas de clasificación. El conjunto TEM contiene más agregados y mejora la clasificación. De hecho, si el conjunto

TEM lo reducimos al tamaño del conjunto SEM, con el método *Random Forest* de 1.000 árboles, la precisión de acierto se reduce de 81,69 % a 78,5 % y el AUC de 0,94 a 0,901. Para el conjunto SEM la mejor precisión de acierto y AUC ha sido mediante el *Random Forest* de 100 árboles, con unos resultados de 79,5 % y 0,931 respectivamente, ligeramente inferior al conjunto TEM reducido. A esto se añade el hecho de que es la técnica de microscopía más utilizada para la tarea de analizar especímenes a escala nanométrica [RS06], por lo que los resultados obtenidos con este conjunto son considerados más relevantes.

Para el conjunto SEM, el algoritmo con mejor resultado, en cuanto al porcentaje de aciertos, es el *Random Forest* de 100 árboles, con un 79,5 %, en cambio, es el de 1.000 árboles el que obtiene el mejor AUC, con un valor de 0,931. Esto ha supuesto una gran mejora con respecto al estudio realizado para la conferencia DEXA [LdRS+10], en la que con el método *Random Forest* de 1.000 árboles, se obtuvieron unos valores de 73,4 % de *accuracy* y un 0,89 de AUC.

El peor algoritmo de aprendizaje automático es el KNN de 1 vecino, tanto en términos de porcentaje de aciertos, con un 66,5 %, como en cuanto al AUC, con un 0,737. Ningún clasificador supera a los cuatro árboles de decisión, y es el SVM con *Kernel Universal Pearson VII* el que obtiene el mejor resultado en cuanto a precisión de acierto del resto de clasificadores. Según el AUC los KNNs de 20 y 25 vecinos se encuentran tras los tres *Random Forest* analizados.

Por otro lado, para el conjunto TEM, se observa una mejora en todos los algoritmos excepto en el árbol de decisión J48. En concreto, los tres *Random Forest* y los tres SVMs obtienen un porcentaje de aciertos mayor al 80 %. Una de las razones principales es el tamaño del conjunto de datos. De hecho, si reducimos el conjunto TEM a 200 instancias, el tamaño del conjunto SEM, observamos una reducción en la efectividad de los algoritmos. Destaca de nuevo el algoritmo *Random Forest* de 1.000 árboles, con un 81,69 % de aciertos y un AUC de 0,94. La versión de 100 árboles y el *Support Vector Machine* con *kernel* polinomial le siguen de cerca, a pesar de que con el primer conjunto el SVM no obtenía resultados especialmente buenos. Según el AUC destacan, con más de un 0,9, los tres *Random Forest* y todos los KNN excepto el de 1 vecino.

El *overfitting* o sobreajuste, ya nombrado en el apartado 2.5, es un problema que hay que tener en cuenta a la hora de realizar tareas de clasificación. Sus dos inconvenientes más importantes son los siguientes. Por una parte, el introducir más parámetros puede hacer que mejore la clasificación de lo conocido y empeore la de lo desconocido. No es nuestro caso, ya que hemos

comprobado a eliminar los parámetros y la clasificación no empeora. Por otra parte, si hay *overfitting* puede ser irreal la calidad obtenida (precisión de acierto, AUC...) del clasificador. Esto se evita mediante la validación cruzada (del inglés, *cross-validation*).

A pesar de obtener la calidad del clasificador mediante validación cruzada, si que hay *overfitting*, es decir, que clasifica mejor la parte del conjunto destinada al entrenamiento que la de validación. Además, el *Random Forest* es conocido por sobreajustarse, especialmente en tareas con datos ruidosos [SLTW04]. Esto es inevitable ya que es lógico que clasifique mejor lo conocido que lo desconocido. Por la naturaleza del *Random Forest* esto es más pronunciado, para el conjunto TEM, que etiqueta lo conocido con un porcentaje de acierto del 99,87 % y un AUC de 1, y lo desconocido con un porcentaje de acierto del 81,69 % y un AUC de 0,94. Con el método *Support Vector Machine* con *kernel* polinomial, pasa de 94,494 % y un AUC de 0,963 a 81,05 % y 0,879. Es decir, el *Random Forest* se sobreajusta más que el SVM, sin embargo le supera ante lo desconocido.

En cambio, al clasificar un conjunto con un modelo entrenado con un conjunto diferente, las diferencias en el método de captura, la iluminación y las condiciones de captura influyen en los resultados. Así, hemos podido ver cómo se sobreajusta al conjunto de datos de entrenamiento al obtener una clasificación peor a la obtenida mediante validación cruzada. Asimismo, se han eliminado atributos con el fin de reducir el sobreajuste y mejorar la clasificación entre diferentes conjuntos, pero los resultados no han mejorado. En el caso en que el tiempo de procesamiento sea una variable crítica, será interesante la reducción de atributos.

En las matrices de confusión hemos identificado los errores de los clasificadores. Los agregados esferoidales, una morfología aparentemente fácil de diferenciar, no son suficientemente bien clasificados en general. Para superar este problema, sería conveniente contar con un mayor número de agregados de este tipo, sin embargo, con los tipos de *carbon black* que se han utilizado para la presente investigación, no ha sido posible por su baja presencia. En el conjunto artificial, los agregados esferoidales suponen un 22 % de las muestras y sólo el 5 % de ellos son mal clasificados.

Otro enfoque utilizado frente a la baja presencia de agregados esferoidales ha sido el de la fusión de las clases esferoidal y elipsoidal, por su gran similitud morfológica y por las propiedades que proporciona al material con el que es mezclado. La justificación para esta decisión viene dada porque las clases esferoidal y elipsoidal se caracterizan por tener un área superficial si-

milar para el mismo volumen en comparación con los agregados lineales y ramificados. Y es precisamente el área superficial lo más característico de la morfología de un nanomaterial a la hora de interactuar con la matriz [Mar04]. La clasificación con esta alternativa mejora ligeramente. Según el algoritmo *Random Forest* de 1.000 árboles, hasta 80,5 % de precisión de acierto y 0,937 de AUC para el conjunto SEM, y 82,97 % de precisión de acierto y 0,944 de AUC para el conjunto TEM. Sin embargo, se ha preferido mantener en los resultados presentados la clasificación de 4 categorías morfológicas establecida en la literatura [MH72, HMSH93, MG99].

Por otro lado, los agregados lineales no alcanzan una tasa satisfactoria, lo que requiere de una búsqueda de características que los diferencien. Gracias a las matrices de confusión se podrá validar que las nuevas características ayuden a discernir especialmente entre los agregados lineales y los ramificados que es la mayor dificultad que encuentran los algoritmos de aprendizaje automático.

Al presentar los tiempos requeridos por los conjuntos SEM, TEM y artificial, con 200, 781 y 1788 instancias respectivamente, ha sido posible por una parte conocer el tiempo requerido por cada clasificador, así como sacar conclusiones sobre la medida en la que influye el tamaño del conjunto en el tiempo requerido por los clasificadores. En general, los KNNs y Bayes son los más rápidos y el *Random Forest* de 1.000 árboles aventaja al resto en gran medida, superando con el conjunto artificial en más de 200 veces el tiempo requerido por el Bayes ingenuo en generar el modelo, y en más de 40 veces para clasificar una instancia.

Para generar cada modelo y validarlo, el tiempo empleado en operaciones de carga de los ficheros y de salvado de los resultados influye en cierta medida, provocando que el incremento del tiempo del conjunto SEM al TEM sea menor a la proporción de incremento de los datos. Al agrandar todavía más el tamaño, en el conjunto artificial, esta influencia es menor y el aumento del tiempo requerido es proporcional a la diferencia de tamaño. Los tiempos requeridos para clasificar una sola instancia son tan pequeños que se ven enormemente influenciados por las operaciones de entrada y salida de los ficheros. Ya que los accesos a ficheros se realizan en bloques, al clasificar más agregados de golpe el tiempo que se precisa por instancia es menor.

Por último, cabe resaltar que la eliminación de atributos no ha conseguido mejorar la clasificación y que en general los resultados se ven menos influidos por la ausencia de las características peor valoradas.

3.6 Sumario

En este capítulo se ha presentado un enfoque no utilizado antes en este campo, que es el de la categorización de agregados mediante algoritmos de *machine learning* y se ha probado que proporciona resultados equiparables a la clasificación manual, con un 81,69 % de *accuracy* y un AUC de 0,94 con el conjunto TEM. Así, esta tarea tiene un alto grado de subjetividad, ya que los agregados con morfologías intermedias son difíciles de clasificar.

Además, a partir de los resultados presentados y de la reducción del conjunto TEM al tamaño de 200 instancias hemos comprobado que es necesario crear conjuntos de datos grandes para permitir entrenar bien el modelo de aprendizaje automático. Además, el bajo número de agregados esferoidales dificulta su adecuada categorización, para lo que se han encontrado dos alternativas. Por un lado se puede buscar un grado de *carbon black* con más agregados esferoidales para entrenar el modelo de aprendizaje automático. Por otro, se pueden fusionar las clases esferoidal y elipsoidal como una alternativa más rápida y menos costosa para disponer de un conjunto más balanceado, ya que para los algoritmos de aprendizaje automático puede suponer un problema el que haya tanta diferencia en la distribución frecuencial de las clases [EHG07].

También se ha comprobado que las condiciones de captura pueden influir en los resultados, por lo que es aconsejable definir unas pautas para el preprocesamiento del nanomaterial, como sonicación y rejillas utilizadas, y parámetros de captura, como iluminación y desenfoque.

A pesar de los esfuerzos realizados en las tareas de selección de atributos, la eliminación de éstos empeora los resultados de clasificación. Incluso utilizando un método de selección de atributos con un clasificador *Random Forest* de 1.000 árboles embebido, la ausencia de los atributos peor valorados empeora la clasificación. Por esto, sólo se eliminarán en caso de existir conflictos con los tiempos de procesamiento. Así, el extraer algunos parámetros tiene un coste más elevado que otros, y a los algoritmos de aprendizaje automático también les influye el número de características empleadas en los tiempos requeridos para generar sus modelos y para clasificar cada instancia.

Por último, gracias a las matrices de confusión hemos localizado un punto importante de trabajo futuro, que es el de investigar en qué se diferencian los agregados lineales, a partir de lo cual se deberá diseñar un atributo que los caracterice, principalmente de los ramificados.

En resumen, hemos presentado la clasificación de agregados de *carbon black* mediante algoritmos de aprendizaje automático, un enfoque no presente en la

literatura para este campo concreto y que permite obtener la distribución de las categorías morfológicas de un grado con la misma precisión de acierto que realizándolo de manera manual por expertos. Además, se han localizado pautas a seguir para el correcto funcionamiento de los algoritmos y trabajo futuro para mejorar los resultados.

«A fuerza de construir bien, se llega a buen arquitecto.»

Aristóteles (384-322 a. C.)

4

Reconstrucción 3D

Los métodos de reconstrucción 3D proporcionan una caracterización más completa de las partículas de refuerzo. De esta forma diversos parámetros que eran estimados mediante el análisis de imágenes bidimensionales, pueden ser medidos para obtener su valor preciso. Además, esto permitirá en un futuro cercano realizar una esqueletización tridimensional, que proporcionará una información mucho más completa que una bidimensional.

El coste de los métodos actuales hace inviable la caracterización de un material por sus requisitos tanto de procesamiento computacional como de asistencia humana. Por esto, en la presente investigación se ha diseñado un método más sencillo que posibilite automatizar y reducir el tiempo de generación de dichos modelos.

El resultado obtenido ha sido asombroso, mediante sólo dos imágenes ortogonales se ha obtenido más información que con una reconstrucción tomográfica tradicional. El progreso alcanzado es incalculable. Sin embargo, es necesario proseguir la investigación para poder implantarlo en un entorno de producción.

En concreto, las contribuciones realizadas en el campo de la reconstrucción tridimensional son las siguientes:

- Proporcionamos un método para la reconstrucción tridimensional de

agregados de *carbon black* a partir de sólo dos imágenes ortogonales.

- Proveemos de un algoritmo de validación de la reconstrucción mediante una tercera imagen situada entre las dos utilizadas para la reconstrucción, es decir, a 45º de ambas.

- Proveemos un método para extraer información del modelo 3D generado.

- Evaluamos la reconstrucción propuesta mediante la reconstrucción tomográfica y la simulación de Montecarlo.

El resto del capítulo queda organizado de la siguiente forma. En la sección 4.1 se da una breve introducción sobre los métodos de reconstrucción tridimensional. Posteriormente, en la sección 4.2 se exponen los fundamentos de la reconstrucción superficial. En la sección 4.3, se describe el proceso de las reconstrucciones tomográficas y su problemática. A continuación, en la sección 4.4 se presentan los métodos basados en Montecarlo para localizar las partículas en una reconstrucción tomográfica. Posteriormente, en la sección 4.5 se detalla el método propuesto en esta investigación y en la sección 4.6 se describe la evaluación empírica llevada a cabo. A continuación, en la sección 4.7 se discuten los resultados y por último, en la sección 4.8 se resumen las contribuciones de este capítulo.

4.1 Introducción

En el campo de la recuperación de información tridimensional, la mayor parte del trabajo se ha dirigido a la estereografía [SS02], es decir, a la reconstrucción de superficies mediante un par de imágenes. Otro enfoque es el de la reconstrucción de la estructura a partir del movimiento registrado en secuencias de imágenes [FP02].

Estos algoritmos generalmente trabajan sólo con descriptores de superficie, los cuales analizan los bordes físicos de los objetos [PdlC07]. A partir de las diferencias geométricas, se busca el mismo punto en las imágenes y se estima la profundidad mediante triangulación [AR05]. Sin embargo, en las imágenes existe además información monocular sobre la profundidad, como variaciones en la textura, gradientes, desenfoque y color entre otros [SCN08]. Para la reconstrucción a partir de pares estereográficos destaca el software comercial MEX [Ali12, TDH11].

Además, la profundidad también puede estimarse a partir de una sola imagen, aunque sigue siendo una tarea difícil, ya que es ambigua en muchos casos. Para superar este problema se han diseñado algoritmos para objetos concretos, como manos y caras [NIK07]. También se han elaborado métodos para detectar formas a partir de las sombras [ZTCS99, MWW02], y a partir de texturas [LG93, MR97, MP90a], pero no dan buenos resultados en superficies con colores y texturas poco uniformes. Otros estudios han utilizado una figura de referencia que tiene que estar presente en la escena junto al objeto a reconstruir [HS05]. Por último, también se ha combinado información global de la imagen con información previa de la escena mediante regiones aleatorias de Markov [SCN08].

4.2 Reconstrucción superficial

La reconstrucción de la superficie de una microestructura se puede realizar a partir de dos imágenes de la muestra con una leve variación en la inclinación del punto de vista [JJ95]. El resultado de esta operación es una nueva imagen denominada mapa de relieve en el que el valor de cada píxel representa la altura de la superficie en dicho punto.

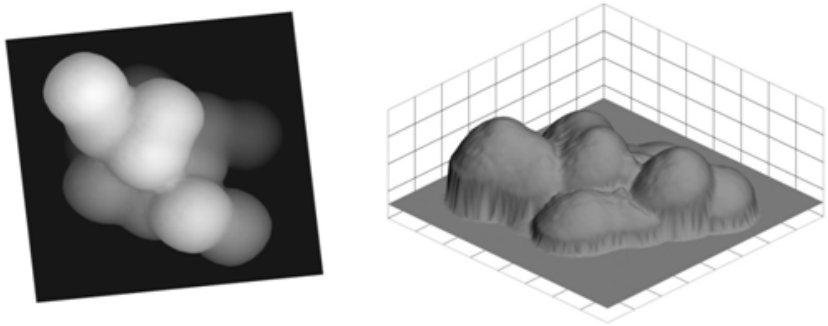


Figura 4.1: Reconstrucción superficial a partir de un mapa de relieve.

La reconstrucción de la superficie se basa en la detección de correspondencias entre las dos imágenes mediante una matriz de disparidad utilizando el análisis bimodal [JJ01]. Así, la obtención del valor (de altura) de un píxel $\boldsymbol{p}$ se puede expresar como:

$$z\left(\mathbf{p}\right) = \frac{-dx\left(\mathbf{p}\right) - cx(\mathbf{p})}{\sin\alpha} \tag{4.1}$$

donde,

$$cx\left(\mathbf{p}\right) = x\left(\mathbf{p}\right)\left(1-\cos\alpha\right) = \left(\mathbf{n}.\mathbf{p}-D\right)\left(1-\cos\alpha\right) \tag{4.2}$$

y

$$dx\left(\mathbf{p}\right) = \mathbf{n}\times\mathbf{d}(\mathbf{p}) \tag{4.3}$$

siendo α el ángulo entre las dos imágenes y $\boldsymbol{n}$ el valor correspondiente en la matriz de disparidad.

La reconstrucción superficial permite obtener información más completa que las imágenes de microscopio, pero debido a las complejas estructuras geométricas que suelen adoptar los agregados de *carbon black*, es preferible trabajar sobre un modelo tridimensional [RS06].

4.3 Reconstrucciones tomográficas

Los métodos tomográficos pueden considerarse como una extensión de la estereoscopía [Jäh02], mediante la cual sólo se obtiene la profundidad de una superficie. En cambio, mediante las tomografías se obtiene la forma tridimensional de los objetos. La tomografía mediante imágenes TEM comenzó a usarse en el campo de la biología estructural a finales de los años 60 [DRK68]. Sin embargo, en el campo de los nanomateriales su uso es relativamente reciente [KKI08]. Los primeros estudios notorios en el área de los polímeros son del 2002 [Wey02, dJK02].

Las tomografías utilizan dispositivos cuya radiación atraviesa un objeto. La radiación empleada es a menudo electromagnética, como en el caso del Microscopio Electrónico de Transmisión (TEM) y de barrido (SEM) o rayos X para la Tomografía Axial Computarizada (TAC). El principio de las tomografías fue postulado por Radon en 1917 [Rad17]. Demostró que la distribución de densidad en un plano puede ser determinada por rayos X tomados en todas las direcciones. Ya que el objeto absorbe parte de la radiación, la pérdida de energía registrada es proporcional a la distancia recorrida por el rayo dentro del objeto. Las imágenes de partida son la proyección en un plano del volumen del espécimen y a su vez, corresponden a un plano en el espacio de Fourier. Por lo que, aplicando la transformada inversa de Fourier a las imágenes, es posible reconstruir el objeto de estudio.

El principal inconveniente de esta técnica es el conocido *cono de sombra* (del inglés, *shadow cone*), que denomina a la región del objeto que no puede ser observada [OGL+10]. En general, la capacidad de rotación de una muestra dentro de un microscopio se limita a los $\pm 70^{o}$ en la mayoría de los casos, pudiendo llegar a los $\pm 90^{o}$. Algunas soluciones para este problema las proporcionan los microscopios que permiten la rotación mediante dos ejes [Mas97] o soportes cilíndricos especiales [KUH+07].

4.4 Algoritmos de Montecarlo

Los algoritmos de Montecarlo [MU49] se basan en la generación de números aleatorios para la obtención de aproximaciones de expresiones matemáticas complejas o la resolución de problemas computacionalmente costosos a base de probar aleatoriamente diferentes soluciones.

En el área de la reconstrucción tomográfica se ha utilizado este algoritmo para simular la disposición de las partículas en un agregado de *carbon black* reconstruido [OGL+10, IIG+11]. Este algoritmo, desarrollado por el equipo EMERG [dPVU12] de la Universidad del País Vasco con el que se ha trabajado en colaboración para la realización de la presente investigación, se basa en un método diseñado por Jinnai [JSK+07], que busca la estructura que mejor encaja en un volumen reconstruido.

El método consiste en cortar el volumen reconstruido en secciones separadas por una distancia uniforme. Cada sección es rellenada por círculos que corresponden a las partículas. Las partículas son consideradas de radio uniforme, sin embargo, los círculos pueden tener un radio inferior ya que la sección puede cortar a las partículas a cualquier altura.

El resultado es extremadamente dependiente del grado de solapamiento permitido. En la literatura el solapamiento de las partículas de un agregado es definido por el parámetro δ [OS97]:

$$\delta = \frac{2a}{l} \tag{4.4}$$

donde a es el radio de la partícula y l es la distancia entre partículas, tomando como puntos de referencia sus centros.

Por otro lado, el grado de solapamiento puede ser definido de manera similar por el parámetro C_{ov} [BFC99]:

$$C_{ov} = \frac{d_p - d_{ij}}{d_p} \tag{4.5}$$

donde d_p es el diámetro de partícula y d_{ij} es la distancia entre dos partículas que se solapan. El parámetro C_{ov} tomo el valor 0 cuando las partículas se tocan sólo en un punto y 1 cuando el solapamiento es total, es decir las dos partículas se encuentran en la misma posición.

Hay que tener en cuenta que estos métodos sólo consideran el solapamiento entre pares de partículas mientras que el método utilizado con el algoritmo basado en Montecarlo contempla el solapamiento de cada partícula con todas las que colisiona [OGL+10, IIG+11].

4.5 Algoritmo RandomSalt

El algoritmo RandomSalt, desarrollado en la presente investigación, obtiene la misma información que el método de Montecarlo que se acaba de describir en la sección 4.4 mediante sólo dos imágenes de partida en este caso, que pueden verse en la Figura 4.2. Además de precisar menos trabajo en la captura de datos también requiere menos trabajo manual en su procesamiento, ya que este procedimiento propuesto es prácticamente automático para la reconstrucción de agregados de *carbon black*.

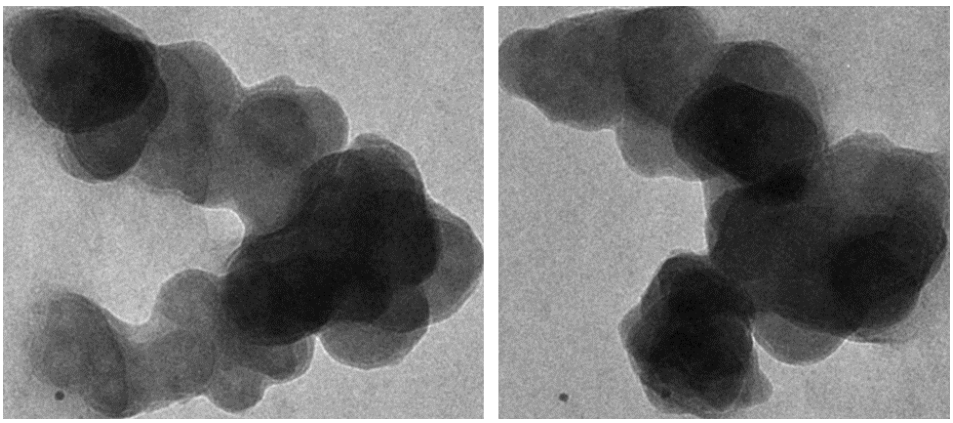


Figura 4.2: Agregado de *carbon black* capturado mediante microscopio TEM a -45º y a +45 º.

Este método se basa en los algoritmos genéticos [Bar54], que aplican una serie de operadores genéticos a un grupo de posibles soluciones, llamadas población, en búsqueda de las mejores en cada iteración o generación. Estos

operadores incluyen diferentes mutaciones a través de las cuales los elementos de la población se van acercando al modelo real.

Posteriormente, a partir de la posición y tamaño de las partículas se realiza un proceso de reconstrucción volumétrica y una generación de su isosuperficie mediante *Marching Cubes* [LC87].

El algoritmo RandomSalt se divide en tres partes principales: el procesamiento de imágenes, el algoritmo genético y la reconstrucción del modelo 3D. De esta forma, el resto de la sección queda organizada como sigue. En el apartado 4.5.1 se explica el procedimiento seguido para segmentar el agregado de las dos imágenes de partida. Posteriormente, en el apartado 4.5.2 se expone como transformar estas dos imágenes en mapas de alturas. A continuación, en la sección 4.5.3 se detalla cómo a partir de los mapas de alturas, mediante un algoritmo genético se van realizando mutaciones para obtener una reconstrucción cada vez más similar al objeto real. Por último, en la sección 4.5.4 se explican los diferentes métodos de visualización del agregado a partir de las posiciones y tamaño de las partículas, y la información que se puede extraer de cada uno de ellos.

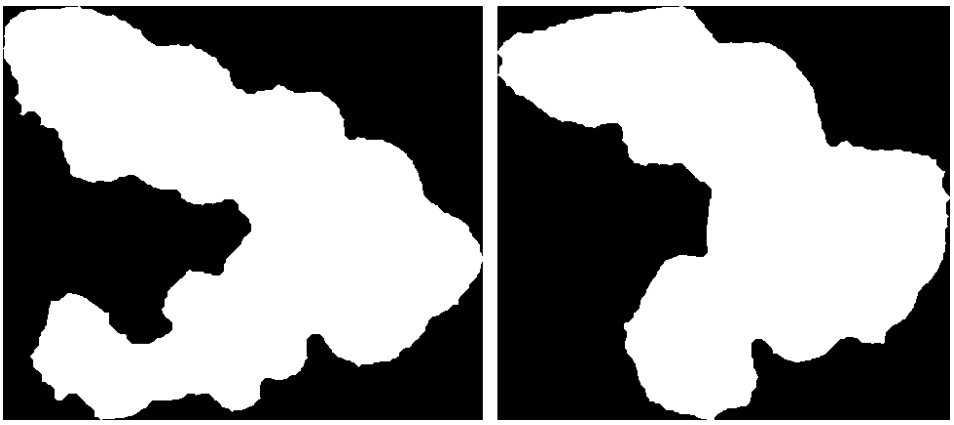


Figura 4.3: Segmentación del agregado de *carbon black* capturado a -45º y a +45º para ser utilizado como máscara.

4.5.1 Segmentación del agregado

El primer paso del algoritmo es la segmentación del agregado capturado en dos imágenes TEM tomadas con una diferencia angular de 90º. El estado de la técnica sobre la segmentación o delimitación de objetos se ha expuesto en el apartado 2.2.5 y posteriormente en el apartado 2.3 al describir los pasos ne-

cesarios para la extracción de características de las imágenes de microscopios electrónicos.

En este caso, el proceso comienza con un suavizado gaussiano, un operador de convolución para eliminar el ruido con el coste de perder un poco de detalle [PdlC07]. Posteriormente, se estima un umbral para discernir entre el agregado y el fondo mediante el método Otsu [Ots75]. A continuación, generamos una imagen binaria considerando que los píxeles por debajo del umbral corresponden al fondo y los que lo superan pertenecen al agregado. Por último, mejoramos la calidad del borde dilatando y erosionando la imagen usando un elemento estructural con forma de disco. Así obtenemos las máscaras, mostradas en la Figura 4.3, que utilizaremos para recortar el agregado y desechar los píxeles que pertenecen al fondo.

4.5.2 Mapa de alturas

Una vez identificado el perímetro del agregado procedemos a la generación del mapa de alturas o de densidad. Éste es la representación de la altura en cada punto de una superficie en una matriz bidimensional. Normalmente se representa mediante una imagen que indica la altura que hay en cada punto a través del color del píxel. En cambio, en nuestro caso, este mapa representa el número de partículas solapadas en cada punto de la imagen y lo generamos de la siguiente forma.

4.5.2.1 Preprocesado de imagen

Primero, se comienza aplicando un filtro bicúbico para reducir el tamaño de la imagen y acortar el tiempo de procesado de los pasos posteriores. Posteriormente, aplicamos un filtro de mediana para reducir el ruido de la imagen, causado inevitablemente por la naturaleza del método de captura. Nótese que el procedimiento seguido es diferente al de segmentación descrito en el apartado 4.5.1, ya que el propósito del procesamiento es diferente en los dos casos. Al segmentar sólo nos interesan los bordes y en este paso sólo nos interesa el interior del agregado.

Por último, invertimos la imagen, obteniendo un fondo predominantemente negro y un agregado cuyos píxeles serán más blancos a mayor cantidad de materia en dicho punto.

4.5.2.2 **Eliminación del ruido ambiente**

El propósito de este paso no es sólo la eliminación del fondo del agregado, sino también del ruido generado por las impurezas presentes en el ambiente en el momento de tomar la imagen mediante un microscopio electrónico. Esto es, el agregado, cuya escala de grises está invertida en este momento, es más blanco que la realidad a causa de estas impurezas.

De este modo, utilizando la imagen segmentada, recortamos el fondo del agregado y calculamos el valor medio de sus píxeles. Así, este valor de la escala de grises es restado en todos los puntos del agregado. En este instante, el fondo es completamente negro, valor 0, y el agregado tiene un valor de intensidad más realista.

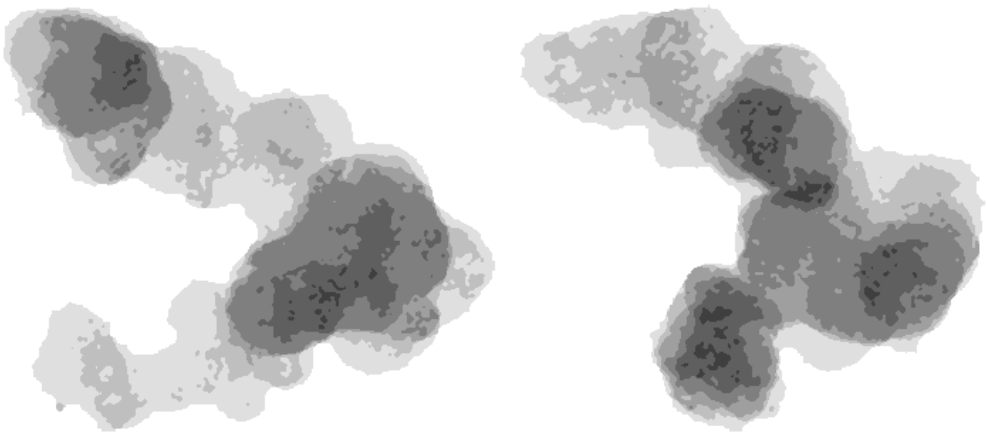


Figura 4.4: Mapa de alturas del agregado de *carbon black* capturado a -45º y a +45º.

4.5.2.3 **Obtención del valor de una partícula**

Para generar el mapa de alturas, es necesario conocer el espesor de una sola partícula. Este dato es obtenido de la siguiente forma. Primero, normalizamos las alturas, convirtiendo la mayor intensidad en blanco puro, valor 255, y la intensidad menor en negro puro, valor 0. De esta forma, conseguimos valores similares de intensidad en las dos imágenes. Esto es necesario, ya que las dos imágenes representan la misma cantidad de materia pero a causa del enfoque automático del microscopio podría parecer que hay más masa en una de las imágenes.

Posteriormente, se divide la escala de grises en intervalos, y se realiza un

recuento del número de píxeles en cada intervalo, discriminando así las alturas que no tienen mucha presencia. Así, el valor intermedio del primer intervalo cuya cantidad de píxeles exceda un umbral definido, será el escogido como la cantidad de materia de una partícula.

4.5.2.4 Generación del mapa de alturas

Una vez que hemos estimado el valor de una partícula, procedemos a la generación del mapa de alturas o de densidad. Para ello dividimos la intensidad de cada punto entre la intensidad de una partícula, obteniendo así una matriz bidimensional con el número de partículas superpuestas en cada píxel.

El cálculo del valor de una partícula depende de la imagen original, de la distribución de las partículas y de la imagen segmentada. Por esta razón, el valor obtenido puede ser demasiado bajo, causando que el mapa de alturas indique que hay un número muy elevado de partículas superpuestas. Para resolver este problema es necesario normalizar la altura del mapa estableciendo un número máximo de partículas superpuestas. Una vez realizado este último paso, obtenemos los mapas de alturas que podemos ver en la Figura 4.4.

4.5.3 Algoritmo genético

El algoritmo diseñado se basa en la teoría de la evolución, pero está simplificado, ya que sólo se aplican mutaciones y no hay operaciones de cruce, es decir, no se mezclan dos soluciones intermedias.

A través de un proceso evolutivo se consigue ir acercándose progresivamente al modelo real. Se busca obtener mapas de alturas similares a los obtenidos a partir de las imágenes TEM a -45º y +45º. Al ser planos ortogonales, la proyección de las partículas en ellos tiene una complejidad computacional muy baja. Así, una partícula que se encuentra en la posición (x, y, z) se proyecta en ambos planos a los puntos (x, y) y (z, y) respectivamente.

El algoritmo comienza inicializando varias soluciones posibles. Posteriormente, hay un proceso iterativo en el que cada una de las soluciones sufren mutaciones. A continuación, las nuevas soluciones son comparadas con las anteriores y se mantienen las mejores. Por último, cuando una solución alcanza la calidad establecida el algoritmo termina.

4.5.3.1 Inicialización

El primer paso del algoritmo consiste en crear una población de 10 instancias o soluciones vacías. Cada una de estas 10 soluciones contendrá una lista de partículas que formarán un agregado.

4.5.3.2 Mutaciones

En cada iteración del proceso evolutivo clonamos todas las soluciones que tenemos y además, las 10 mejores las clonamos de nuevo para potenciar las que destacan y hacer más rápido el proceso.

Seguidamente, aplicamos una mutación aleatoria a cada una de las soluciones clonadas. Las 5 mutaciones posibles son las siguientes:

1. Añadir partícula: esta mutación genera un punto tridimensional aleatoriamente y comprueba que su proyección en los planos de las dos imágenes originales esté dentro de las dos máscaras. El radio de la partícula es generado aleatoriamente entre unos valores máximo y mínimo establecidos manualmente al comienzo del proceso.

2. Eliminar partícula: esta mutación elige una partícula aleatoriamente para su eliminación.

3. Mover partícula: esta mutación elige una partícula aleatoriamente y genera una posición aleatoria a la que trasladar la partícula. Al igual que al añadir una partícula, su posición central debe poder proyectarse dentro de las dos máscaras.

4. Mover partícula un poco: esta mutación elige una partícula aleatoriamente y escoge una dirección aleatoria entre norte, sur, este u oeste, para mover la partícula un píxel. Al igual que al añadir una partícula, su posición central debe poder proyectarse dentro de las dos máscaras. Esta mutación es especialmente útil cuando la reconstrucción se encuentra en un estado avanzado y el añadir una partícula o insertarla en la mayoría de los sitios es perjudicial.

5. Redimensionar partícula: esta mutación elige una partícula aleatoriamente para cambiar su tamaño. El nuevo radio de la partícula es generado aleatoriamente entre los valores máximo y mínimo establecidos.

Además, al aplicar una mutación se comprueba que la esfera se encuentre completamente dentro del prisma cuadrado creado por las dos imágenes.

4.5.3.3 Selección

Para determinar las mejores instancias hemos creado una función de ajuste o de idoneidad (del inglés, *fitness*). Esta función calcula las diferencias entre los mapas de alturas obtenidos de las dos imágenes originales y las proyecciones correspondientes creadas a partir de las partículas de la solución, que pueden observarse en la Figura 4.5.

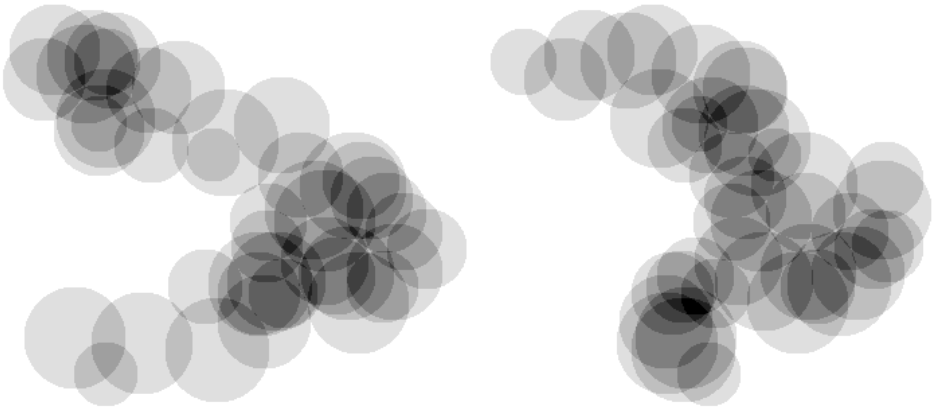


Figura 4.5: Proyecciones de las partículas a los planos correspondientes de las imágenes originales.

Además, una instancia es penalizada por la porción de máscara que no es rellenada por partículas. Adicionalmente, las partículas aisladas reducen la calificación obtenida por la función de *fitness* proporcionalmente al radio y a la distancia de la partícula más cercana.

En concreto, la función de *fitness* diseñada se muestra en la siguiente ecuación:

$$Fitness = 1 - \frac{diferencias + penalizacion}{diferencias\ máximas} \tag{4.6}$$

donde las $diferencias$ se calculan según la siguiente ecuación:

$$\begin{aligned} diferencias = & \sum_{x=1}^{x=X} \sum_{y=1}^{y=Y} |proyecciónIzq(x,y) - originalIzq(x,y)| \\ & + \sum_{z=1}^{z=Z} \sum_{y=1}^{y=Y} |proyecciónDcha(z,y) - originalDcha(z,y)| \end{aligned} \tag{4.7}$$

donde la coordenada x puede tomar el valor máximo de X, el ancho del agregado visto a -45º, la coordenada y puede tomar el valor máximo de Y, la altura del agregado visto a -45º, $proyecciónIzq$ es el mapa de alturas generado tras proyectar las partículas de la solución que se está evaluando al plano a -45º, $originalIzq$ es el mapa de alturas obtenido de la imagen original correspondiente a dicho plano, $proyecciónDcha$ y $originalDcha$ son los mapas de alturas correspondientes al plano a +45º.

Para calcular la *penalización* por no rellenar el área de la máscara del agregado original del plano a -45º y por partículas sueltas se sigue el Algoritmo 1. Éste describe como se recorren los mapas de alturas generados proyectando las partículas de la solución (ver Figura 4.5) y se comparan punto a punto con las máscaras obtenidas de las imágenes originales (ver Figura 4.3). En caso de encontrar un píxel marcado en la máscara como región, y que tiene un valor nulo en el mapa de alturas, se aumenta la penalización aplicada a la función de *fitness*. Además, se recorre cada una de las partículas en búsqueda de su partícula más cercana. En el caso de que con ésta no supere la distancia mínima establecida de 20 *vóxeles* entre sus extremos, se aplica una penalización proporcional a la separación y al tamaño de la partícula.

Por último, para calcular las $diferencias\ máximas$, se aplica la siguiente ecuación:

$$diferencias\ máximas = anchura \cdot altura \cdot profundidad \cdot 4 \tag{4.8}$$

De esta forma se obtiene un número de diferencias que no será superado por las diferencias de los mapas de alturas junto con las penalizaciones, lo que provocaría un *fitness* negativo.

4.5.3.4 **Finalización**

Para marcar el fin del proceso se establece el *fitness* deseado. Adicionalmente, se puede establecer un tiempo o número de iteraciones máximo y por último es posible detener el algoritmo manualmente si se observa que la solución alcanzada es aceptable.

Algoritmo 1: Algoritmo para calcular las penalizaciones.

Entrada: mapas de alturas generados proyectando las partículas sobre los planos a -45º y +45º, *proyecciónIzq* y *proyecciónDcha*, sus respectivos mapas de alturas originales *originalIzq*, *originalDcha* y lista de partículas, *partículas*.
Salida: suma de las penalizaciones, *penalización*.

```
Penalización por no rellenar parte de la máscara.
for x ← 1 to anchura(proyecciónIzq) do
    for y ← 1 to altura(proyecciónIzq) do
        if proyecciónIzq(x,y) = 0 and originalIzq(x,y) ≠ 0 then
            penalización ← penalización + 2
        end
    end
end
for z ← 1 to anchura(proyecciónDcha) do
    for y ← 1 to altura(proyecciónDcha) do
        if proyecciónDcha(z,y) = 0 and originalDcha(z,y) ≠ 0 then
            penalización ← penalización + 2
        end
    end
end
Penalización por partículas sueltas.
for i ← 1 to tamaño(partículas) do
    for j ← i + 1 to tamaño(partículas) do
        distancia ← CalcularDistancia(partículas(i), partículas(j))
        distanciaMin ← radio(partículas(i)) + radio(partículas(j)) - 20
        if distancia − distanciaMin < separacionMinEncontrada
        then
            separacionMinEncontrada ← distancia - distanciaMin
        end
    end
    if separacionMinimaEncontrada > 0 then
        penalización ← penalización + separacionMinEncontrada ·
        área(partícula(i))²
    end
end
```

4.5.4 Reconstrucción tridimensional y visualización

A partir de la información obtenida, es decir, de la posición y tamaño de las partículas, recreamos el agregado de *carbon black* original en un espacio tridimensional. El modelo resultante, además de utilizarse para apreciar visualmente la morfología del agregado, incluye información volumétrica y superficial adicional.

En la aplicación desarrollada proporcionamos al usuario 3 modos de representación diferentes. Cada uno de ellos permite que se observe el agregado de manera distinta y a su vez permite el cálculo de nueva información.

- Representación basada en esferas: es la más simple de las tres. Utiliza esferas para representar cada partícula del agregado. Proporciona un método rápido y simple para el hardware gráfico de dibujar la superficie del agregado. Así, el usuario puede comprobar con esta representación si la mejor solución actual es suficientemente buena.

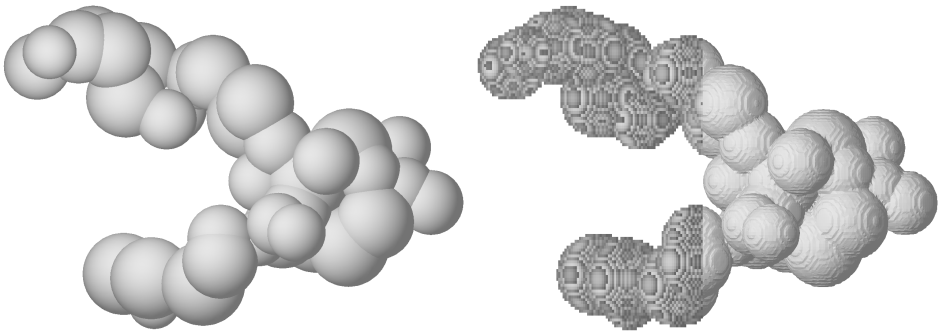


Figura 4.6: Representación basada en esferas y representación volumétrica fusionada con la representación de *isosuperficie*.

- Representación volumétrica: en este método de representación sólo se muestra la información volumétrica. Para ello se crea una red de *voxelización* (del inglés, *voxelization grid*) cuyos elementos individuales, llamados *vóxeles*[1], tienen un valor diferente dependiendo de la distancia al punto exterior más cercano. Este valor es convertido a una *paleta* de colores interpolable y dibujado en la pantalla. Además, se permite al usuario cortar el agregado para visualizar su interior.

[1]Un *vóxel* es un elemento volumétrico que supone la unidad mínima procesable de una matriz tridimensional. Es análogo al píxel de una imagen bidimensional.

- Representación de *isosuperficie*: tomando como partida la red de *voxelización* de la representación volumétrica, se crea una malla de triangulación mediante nuestra implementación mejorada del algoritmo *Marching Cubes* [LC87]. Gracias a esto conseguimos una superficie que se asemeja en mayor medida a la original.

En la Figura 4.6 se puede observar por un lado la representación basada en esferas y por otro lado una fusión de la representación volumétrica con la representación de *isosuperficie*.

4.6 Evaluación empírica

Para la evaluación del método propuesto hemos capturado un agregado de *carbon black* con un Microscopio Electrónico de Transmisión JEM-2200FS/CR (JEOL, Japan) equipado con un escáner ULTRASCAN 4000 SP (4008 x 4008 píxeles) y una cámara CCD lenta con capacidad de enfriado (Gatan, UK). El grado de *carbon black* utilizado ha sido el CSX 691, suministrado por Cabot Corporation. Para analizar los especímenes hemos seguido el procedimiento ASTM D3849-07 [Ame07] y de las imágenes capturadas, para este algoritmo hemos utilizado las correspondientes a los ángulos -45º, 0º y +45º.

En este apartado se busca responder a las siguientes preguntas:

- *¿Cuánto difieren las características extraídas del agregado reconstruido con respecto a la reconstrucción tomográfica?*
- *¿Cuánto difieren las características extraídas del agregado reconstruido con respecto a la simulación de Montecarlo?*
- *¿Es posible sustituir el método tradicional de las reconstrucciones tomográficas mediante nuestro algoritmo propuesto?*
- *¿Cuánto tiempo requiere el algoritmo para obtener una solución satisfactoria y cuántas iteraciones supone?*

Las preguntas primera y segunda hacen referencia a los valores calculados de área superficial y volumen. Así, en la discusión de resultados se compararán con los valores que proporcionan la reconstrucción tomográfica y la simulación basada en Montecarlo. Además, con este último método también se comparará el número de partículas estimado. La tercera pregunta depende de las dos

preguntas anteriores, del *fitness* obtenido por la mejor solución y del *fitness* de validación con una imagen intermedia a las dos utilizadas para generar el modelo. La última pregunta es una cuestión de eficiencia y su interés reside en que el objetivo principal de este algoritmo es el análisis de un conjunto de datos de gran tamaño para el estudio de sus características.

Como se ha descrito previamente, hemos procesado las dos imágenes ortogonales para obtener sus mapas de alturas, determinando una altura máxima de 6 partículas. Después, hemos creado una población de 10 elementos o individuos vacíos. Cada uno de ellos contenía una lista de partículas con sus posiciones y tamaños. El tamaño de las partículas que hemos establecido varía entre 19 y 38 nm. Además, cada individuo contenía dos proyecciones que se actualizaban conforme a los cambios que sucedían en la lista de partículas.

Una vez inicializados los 10 individuos, los hemos clonados dos veces. A cada uno de estos 20 clones les hemos aplicado una mutación aleatoria de las cinco disponibles: añadir, eliminar, mover, mover un poco y redimensionar. En este momento se alcanza la población máxima de 30 elementos que hemos mantenido al finalizar cada iteración. En este punto hemos evaluado y ordenado los individuos, considerándo la mejor solución como el progreso alcanzado por el algoritmo.

La forma de evaluar cada individuo y poder distinguir que solución intermedia es mejor que otra ha sido mediante la función de ajuste o función de *fitness*. Ésta consiste en comparar los dos mapas de alturas generados a partir de las imágenes originales con los mapas de alturas de cada elemento de la población generados al proyectar sus partículas sobre los mismos planos ortogonales. Adicionalmente, la función de *fitness*, tiene en cuenta las porciones de máscara que quedan sin rellenar y las penaliza para favorecer a los individuos que distribuyen mejor las partículas en las dos proyecciones generadas por el volumen original. Por último, también penaliza la existencia de partículas aisladas proporcionalmente al tamaño de la partícula y a la distancia de su partícula más cercana.

Durante el resto del proceso, cada iteración del algoritmo comienza con 30 individuos que son clonados para aplicarles mutaciones. Además, hemos clonado los 10 mejores una vez más para favorecer la evolución de los más *fuertes* y disminuir el número de iteraciones necesarias para encontrar una solución satisfactoria. Una vez hemos aplicado las mutaciones a los 40 clones, éstos han sido comparados con los 30 individuos que había al comienzo de la iteración y se han guardado sólo los 30 mejores elementos, que han pasado a la siguiente iteración. El tamaño de la población ha sido elegido empíricamente

tras utilizar diferentes configuraciones, para encontrar un compromiso entre la efectividad y la eficiencia. Una población grande puede evitar la aparición de un máximo local, es decir, que el algoritmo se estanque en una solución no óptima. Por otro lado, dicha población supone una mayor carga computacional.

El proceso puede llegar a su fin por 4 causas posibles: cuando lleva ejecutándose un tiempo definido, tras un número determinado de iteraciones, al alcanzar el *fitness* deseado o cuando manualmente se observa que la reconstrucción es satisfactoria. En este caso, se ha establecido un tiempo máximo de 1 hora, tras el que manualmente se ha comprobado la calidad de la reconstrucción. La solución obtenida ha contenido 42 partículas.

Además, para la evaluación del resultado hemos empleado 4 cálculos. Primero, el *fitness* obtenido es un indicador de la similitud de las proyecciones. Posteriormente, se ha validado con una tercera imagen de microscopio. Además, se ha comparado el volumen y el área superficial con los de la reconstrucción tomográfica y la simulación de Montecarlo.

Primero, el *fitness* que hemos obtenido ha sido de 0,9995301, lo que demuestra que los mapas de alturas generados por las dos imágenes ortogonales originales a -45º y a +45º tienen una gran similitud con las generadas al proyectar las partículas de la mejor solución sobre los mismos planos. En la Figura 4.7 se muestra la evolución del *fitness* a lo largo de las iteraciones. Para las 62.390 iteraciones mostradas en la gráfica, el algoritmo se ejecutó durante 18 horas y 22 minutos. Al cabo de 1 hora se habían realizado 3397 iteraciones, y la calidad de la reconstrucción ya era lo suficientemente buena. El valor máximo se alcanzó a las 5 horas y 53 minutos, en la iteración 20.025.

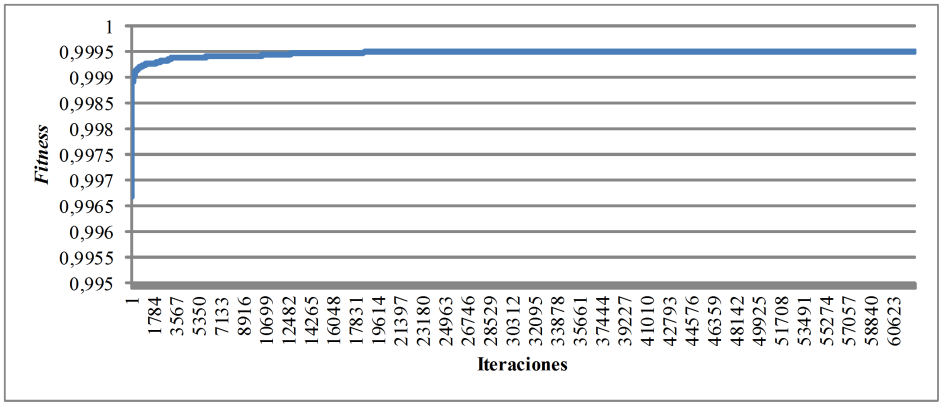


Figura 4.7: Evolución del *fitness* con respecto a las iteraciones.

Posteriormente, para validar el modelo 3D, hemos utilizado una tercera imagen de microscopio capturada a 0º, o dicho de otro modo a 45º de los dos planos ortogonales. Dicha imagen se muestra en la Figura 4.8 de la que se incluye también su mapa de alturas y el mapa generado proyectándolo a dicho plano.

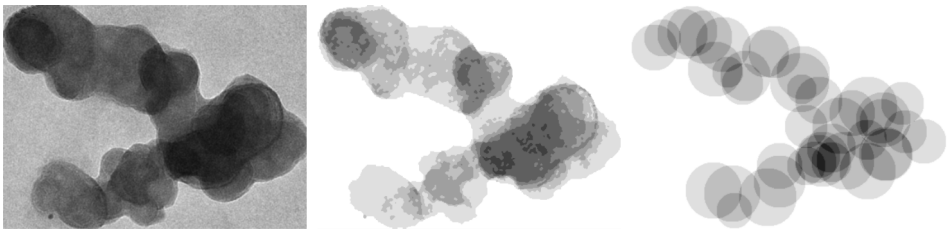


Figura 4.8: Imagen del agregado tomada a 0º, de la que se muestra también su mapa de alturas y el mapa generado proyectándolo a dicho plano.

Esta tercera imagen no se ha utilizado para elegir qué soluciones eran mejores que otras, sino sólo para validar el modelo obtenido una vez terminado el proceso. Para este paso hemos realizado el mismo procesado descrito en el apartado 4.5.2 para obtener el mapa de alturas. Posteriormente, para proyectar las esferas hemos aplicado una transformación de coordenadas por rotación mediante la siguiente ecuación:

$$x'_A = x_A \cos\alpha + z_A \sin\alpha \tag{4.9}$$

donde x'_A es la nueva coordenada en el eje de las abscisas que se quiere obtener para el punto A, x_A es la coordenada x original para el punto A, α es el ángulo de rotación, en este caso 45º y z_A es la coordenada z original para el punto A. En cambio $y'_A = y_A$ ya que en el eje de las ordenadas no se aplica ninguna rotación. Por tanto, dicha coordenada coincide tanto para los dos planos ortogonales utilizados por el algoritmo genético, como para el plano de validación. Así, el *fitness* que hemos obtenido con esta proyección ha sido de 0,9978869, lo que demuestra que a pesar de que la imagen a 0º no ha sido utilizada para la generación del modelo 3D, éste tiene una proyección en el mismo plano muy similar al mapa de alturas correspondiente a dicha imagen.

Por último, se han obtenido los valores de volumen y área superficial. Para ello, se ha realizado primero una representación volumétrica a partir de las posiciones de las partículas y de su tamaño. La red de *voxelización* creada ha dado lugar a un volumen de $1{,}21\times10^7$ nm^3. Seguidamente, a partir de esta representación se ha generado su isosuperficie mediante una malla de

triangulación proporcionando un área superficial de $4{,}45\times10^5$ nm^2.

4.7 Discusión de los resultados

Los resultados demuestran que una nueva forma de reconstruir agregados es posible y que gracias a ella se podrán caracterizar grandes conjuntos de agregados de una forma mucho más precisa y completa que en la actualidad.

Así, los valores obtenidos son similares a los que proporciona la reconstrucción tomográfica. Para su obtención, el equipo EMERG [dPVU12], capturó el mismo agregado de *carbon black* de -60º a +60º a intervalos de 1,5º. A partir de estas 81 imágenes se obtuvieron unos valores de área superficial y volumen de $5{,}93\times10^5$ nm^2 y $1{,}69\times10^7$ nm^3 respectivamente.

Por otro lado, de nuevo el equipo EMERG [dPVU12], rellenó de partículas la reconstrucción tomográfica mediante una simulación de Montecarlo. Dicho algoritmo introdujo en dicho volumen 105 partículas con un radio uniforme de 35,9 nm, lo que proporcionó unos valores de área superficial y volumen de $7{,}459\times10^5$ nm^2 y $1{,}468\times10^7$ nm^3.

En cambio, para el mismo agregado y con sólo 2 imágenes, según la representación volumétrica e isométrica del modelo obtenido mediante el algoritmo genético RandomSalt, el área superficial y el volumen han sido de $4{,}45\times10^5$ nm^2 y $1{,}21\times10^7$ nm^3 respectivamente. Además, mediante este método se han obtenido 42 partículas. Esta cantidad difiere en gran medida de las obtenidas a partir del algoritmo de Montecarlo a pesar de no haber la misma proporción en cuanto a la diferencia de área superficial y de volumen. Esto se debe al diferente grado de solapamiento existente entre los dos métodos.

Hay que tener en cuenta que con las tomografías, al no utilizar la información de profundidad y aprovechar sólo la silueta del agregado, hay una parte del agregado que no es posible observar, el llamado *cono de sombra* [OGL+10]. Además, hay que realizar un segmentado de los planos ortogonales (del inglés, *orthoslices*) obtenidos después de la reconstrucción. Esta intervención manual introduce errores adicionales al proceso de reconstrucción en si mismo [Fra06].

Por tanto, los valores de la reconstrucción tomográfica, no pueden tomarse como datos exactos sino como información de referencia. Del mismo modo, el algoritmo basado en Montecarlo, parte de una solución no perfecta a la que añade un error adicional, con la limitación añadida de que el tamaño de partícula es uniforme. Además, aunque añade información, como es la localización de las partículas, requiere de una reconstrucción tomográfica como

dato de entrada, por lo que no es factible para su aplicación en un conjunto elevado de agregados por la cantidad de trabajo manual que supone.

Por último, el *fitness* de la mejor solución, obtenido al comparar los mapas de alturas de las imágenes originales ortogonales con las proyecciones de la lista de partículas ha sido de 0,9995301. Asimismo, el *fitness* de validación también es excelente, dando un 0,9978869 para la comparación de la proyección de la lista de partículas con el mapa de alturas de la imagen original capturada a 0º de inclinación, la cual ha sido desconocida para el algoritmo hasta el momento de la validación.

4.8 Sumario

Para generar un modelo 3D normalmente se capturan alrededor de 100 imágenes para realizar con ellas una reconstrucción tomográfica. En cambio, nosotros hemos desarrollado otra alternativa realmente innovadora, con sólo 2 imágenes ortogonales, a -45º y + 45º y un algoritmo genético.

Estas imágenes, que están en escala de grises, se transforman en mapas de densidad teniendo en cuenta que el nivel de gris está relacionado con el número atómico, la densidad y el espesor del objeto analizado.

El algoritmo comienza creando varios modelos inicialmente vacíos que se mantienen en paralelo. En cada iteración, se clonan las soluciones y se les aplican mutaciones aleatorias. Estas mutaciones son añadir, eliminar, mover y cambiar de tamaño una partícula. Las mejores soluciones se mantienen para la próxima iteración.

Cada solución contiene una lista de partículas que es proyectada como se muestra en la Figura 4.9 generando dos mapas de alturas o de densidad. Una de las partículas está destacada y se indica cómo se proyecta sobre los dos planos ortogonales.

Para determinar la bondad de una solución se utiliza una función de *fitness* que se basa en tres conceptos.

- Se miden las diferencias entre las proyecciones de las soluciones (ver Figura 4.9) y los mapas de densidad de las imágenes originales (ver Figura 4.4).

- La porción de región que queda sin rellenar es penalizada otra vez más.

- Se penalizan las partículas desagregadas ya que queremos reconstruir un objeto conexo.

En una hora se obtiene un resultado adecuado y del modelo 3D se pueden extraer sus características morfológicas.

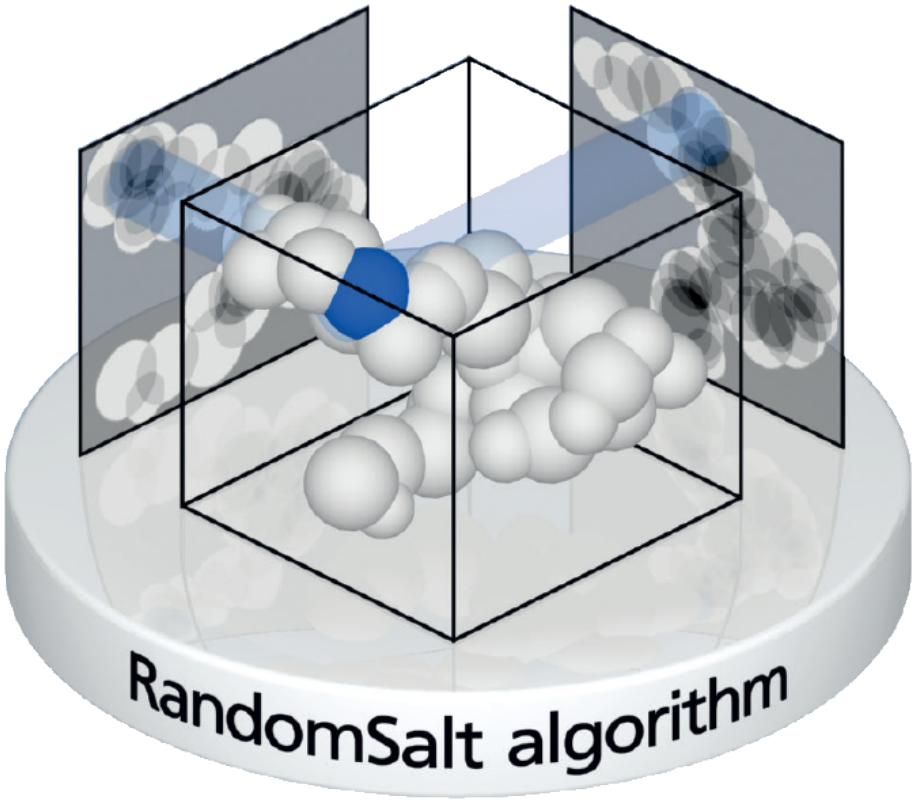


Figura 4.9: Modelo 3D de un agregado de *carbon black* con sus proyecciones.

Gracias al método aquí propuesto se puede llegar a realizar una caracterización automática más completa de los agregados, actualmente limitada al análisis bidimensional para grandes conjuntos de datos, ya que el coste de las reconstrucciones tomográficas es inviable. Aunque existen simulaciones en la literatura, su precisión es limitada y están más orientados a la distribución del material de relleno que a la caracterización precisa de agregados individuales [MLH12].

«El futuro es maravilloso pero hay que hacerlo.»

Elisa Carrió (1956-presente)

5

Conclusiones

Tras haber expuesto el proceso de investigación seguido, las contribuciones aportadas y la evaluación de las mismas, es el momento de revisar la hipótesis y los objetivos iniciales y conocer su grado de consecución. Además, se expondrán las limitaciones del sistema propuesto y se dará a conocer el trabajo futuro identificado para los tres pilares de esta investigación sobre el *carbon black*: caracterización, categorización y reconstrucción 3D.

Así, el capítulo se organiza de la siguiente forma. La sección 5.1 presenta las diferentes validaciones de los experimentos realizados. A continuación, la sección 5.2 resume los resultados obtenidos comparándolos con la hipótesis de partida y los objetivos que surgen de ella. Posteriormente, en la sección 5.3 se exponen las aplicaciones identificadas para las que las contribuciones supondrán una mejora competitiva. Seguidamente, en la sección 5.4 se desvelan las limitaciones encontradas. Además, en la sección 5.5 se aportan soluciones a dichas limitaciones y se marcan las líneas futuras de la investigación. Por último, en la sección 5.6 se concluye esta disertación con las consideraciones finales del autor.

5.1 Síntesis de la validación del sistema

En el presente trabajo de investigación se han llevado a cabo varias validaciones empíricas con el objetivo de comprobar cada uno de los componentes expuestos. A continuación, se resumen las conclusiones más destacadas.

En primer lugar, con el fin de validar las nuevas características que hemos diseñado para extraer información de las imágenes de microscopios de agregados de *carbon black*, empleamos cuatro métodos de selección de atributos, dos univariables y dos multivariables. Además, se han aplicado sobre tres conjuntos de datos, uno de imágenes SEM (Microscopio Electrónico de Barrido), otro de imágenes TEM (Microscopio Electrónico de Transmisión) y otro de imágenes artificiales generado para tener un conjunto con las clases balanceadas. Tanto el conjunto SEM, como el TEM no están balanceados, lo que puede ser un problema con los métodos tradicionales de aprendizaje automático [EHG07]. Aun así, cabe resaltar que el conjunto TEM es el que consideramos como el más relevante, tanto por su tamaño como por su diversidad en las condiciones de captura. Asimismo, es la técnica de microscopía más utilizada para la tarea de analizar especímenes a escala nanométrica [RS06].

Para evaluar lo relacionado que está cada atributo con la forma, hemos etiquetado manualmente los agregados en las cuatro categorías morfológicas definidas por Herd [HMH92]. Cabe mencionar que esta tarea no resulta trivial, ya que las diferencias entre una clase y otra no están claramente delimitadas y resulta un tanto subjetiva. Además, hemos utilizado dos métodos univariados: Chi-cuadrado y Relación de Ganancia de Información. El mejor atributo para los dos métodos, *relaciónÁreaÁreaConvexa*, es un parámetro comúnmente usado en la literatura [PI97, Rus07]. Según el método Chi-cuadrado se encuentra a continuación el atributo *áreaÁreaBoundingBox*, utilizado en la literatura para el tratamiento de imagen pero no en concreto para el *carbon black*.

Además, los resultados han demostrado satisfactoriamente la bondad de nuestras características, superando algunas a las existentes en la literatura. Esto es especialmente cierto según los dos métodos multivariados: *ReliefF* y un método *embebido* con el clasificador *Random Forest* como medida de utilidad.

En concreto, *porcentajeSegmentos* y *porcentajeCortesSegmentos* son los dos mejores parámetros, superando a todos los existentes en la literatura para el conjunto TEM según el método de selección de atributos basado en *Random Forest*, el método más relevante para nosotros, ya que es el que ha obtenido mejores resultados de clasificación.

En general, *mediaÁreaÁreaConvexa* y *cambioÁreaSuavizada*, también atributos propios, han resultado relevantes en la mayoría de los casos. Posteriormente, con un resultado todavía aceptable, se encuentran los parámetros *desviaciónÁreaÁreaConvexa, mediaÁreaÁreaTriángulo* y por último *desviaciónÁreaÁreaTriángulo*. Además, 6 de los parámetros propios han obtenido una relevancia casi nula. Éstos han sido *sectoresVacíos, cambiosDerivada* y los relacionados con las ramas del agregado: *desviaciónRamas, medianaRamas, mediaRecortadaRamas* y *mediaWindsorizadaRamas*.

Estos cuatro últimos parámetros propios dependen de un proceso de esqueletización que es necesario mejorar. Esta técnica ha sido utilizado con éxito en la literatura [MG99] y es claro que de este tipo de características se puede extraer información morfológica.

En cuanto a los tiempos de procesamiento, el algoritmo de selección de atributos mediante el *Random Forest*, supera al resto de una manera desproporcionada, especialmente para el de 1.000 árboles. Sin embargo, siguen siendo más que aceptable, teniendo en cuenta que es una tarea que normalmente no requiere demasiada urgencia, como puede ser la reducción de la dimensionalidad de un conjunto para la creación de un modelo de clasificación. Sin embargo, es posible que en ciertos casos se prefiera reducir ligeramente la calidad a cambio de disminuir notablemente el tiempo, como puede ser con grandes conjuntos de datos o en caso de querer encadenar el proceso de selección de atributos con la generación de un modelo de clasificación frecuentemente actualizado.

Por otro lado, hemos llevado a cabo una categorización automática de los agregados de *carbon black* en los cuatro tipos morfológicas sobre los tres mismos conjuntos de datos mencionados para la etapa de caracterización. Esta tarea la hemos realizado mediante cuatro tipos de clasificadores, con diferentes *kernels* y configuraciones, formando un total de 17 algoritmos. La evaluación de los resultados se ha llevado a cabo en términos de *accuracy*, o precisión de acierto y AUC (Área por debajo de la curva ROC). Esta última, no se ve afectada ante un cambio en la proporción de las instancias de cada clase, a diferencia de otras métricas como el porcentaje de aciertos [Faw06].

Ya que en la literatura no existen medidas sobre la calidad de las clasificaciones y que es una tarea con un alto grado de subjetividad, hemos llevado a cabo clasificaciones independientes por 6 expertos y se ha realizado una clasificación por voto. Posteriormente, hemos comparado cada clasificación individual con la efectuada por voto y la media obtenida ha sido de un 80 % de precisión de acierto. Este valor se ha considerado como el umbral necesario

para categorizar de manera automática.

Para el conjunto SEM, el algoritmo con mejor resultado en cuanto al porcentaje de aciertos, es el *Random Forest* de 100 árboles, con un 79,5 %, en cambio, es el de 1.000 árboles el que obtiene el mejor AUC, con un valor de 0,931. El peor algoritmo de aprendizaje automático es el KNN de 1 vecino, seguido por el Bayes ingenuo. Ningún clasificador supera a los cuatro árboles de decisión, y es el *Support Vector Machine* con *Kernel Universal Pearson VII* el que obtiene el mejor resultado en cuanto a precisión de acierto del resto de clasificadores. Según el AUC, los KNNs de 20 y 25 vecinos se encuentran tras los tres *Random Forest* analizados.

Por otro lado, para el conjunto TEM, se observa una mejora en todos los algoritmos excepto en el árbol de decisión J48. En concreto, los tres *Random Forest* y los tres SVMs obtienen un porcentaje de aciertos mayor al 80 %. Una de las razones principales es el tamaño del conjunto de datos. Si reducimos el conjunto TEM a 200 instancias, el tamaño del conjunto SEM, observamos una reducción en la efectividad de los algoritmos. Destaca de nuevo el algoritmo *Random Forest* de 1.000 árboles, con un 81,69 % de aciertos y un AUC de 0,94. La versión de 100 árboles y el SVM con *kernel* polinomial le siguen de cerca, a pesar de que con el primer conjunto el SVM no obtenía resultados especialmente buenos. Según el AUC destacan, con más de un 0,9, los tres *Random Forest* y todos los KNN excepto el de 1 vecino.

En cambio, para el conjunto artificial, ha destacado el SVM con *Kernel Universal Pearson VII*, con el que se ha alcanzado un 85,07 % de *accuracy*. Según el AUC, vuelve a destacar el *Random Forest* de 1.000 árboles. Mediante las matrices de confusión, vemos como en este conjunto, los agregados esferoidales sí son clasificados correctamente, con sólo el 5 % de ellos mal identificados. Así, el que las categorías del conjunto están balanceadas adecuadamente ayuda a mejorar la categorización de esta clase, que es minoritaria en nuestros otros dos conjuntos de datos. Además, gracias a las matrices de confusión hemos podido observar en todos los conjuntos la dificultad para clasificar a los agregados lineales.

En cuanto a los tiempos de procesamiento, los KNNs y Bayes son los más rápidos y el *Random Forest* de 1.000 árboles supera al resto en gran medida, superando con el conjunto artificial en más de 200 veces el tiempo requerido por el Bayes ingenuo en generar el modelo, y en más de 40 veces para clasificar una instancia. Aun así, los tiempos no son excesivos y no deberían suponer ningún problema, pero en el caso contrario, aconsejamos la reducción del número de árboles o de las características menos relevantes.

La última contribución, la reconstrucción tridimensional mediante un algoritmo genético ha sido validada con el método clásico de la reconstrucción tomográfica y mediante una simulación basada en Montecarlo. Así, el método diseñado ha obtenido unos datos satisfactorios en mucho menos tiempo, con mucho menos trabajo manual y con sólo dos imágenes ortogonales.

En sólo una hora, hemos conseguido buenos resultados, y con más tiempo la reconstrucción sigue mejorando. Además, hemos diseñado un método de validación con una imagen adicional que no ha sido usada para la reconstrucción del agregado.

5.2 Resumen de los resultados obtenidos

En esta sección se resumen los resultados obtenidos a lo largo de la presente investigación para la validación de la hipótesis de partida. Para recordarla, esta enuncia lo siguiente:

> *Es posible desarrollar un sistema automatizado para el control de la calidad de nanomateriales formados por negro de humo y para la mejora de su proceso productivo, por medio del tratamiento de imágenes de microscopio a través de su caracterización y posterior aplicación de técnicas de machine learning.*

A continuación, se resumen las contribuciones en el área de la caracterización, categorización y reconstrucción 3D de agregados de *carbon black*, cuyas validaciones han sido recopiladas en la sección 5.1. Gracias a ellas posibilitamos el análisis automatizado de su morfología, tarea de gran utilidad tanto para el control de calidad como para la fabricación de nuevos materiales con propiedades mejoradas.

1. **Un método para la extracción de características de nanoagregados.** Hemos propuesto una técnica para el tratamiento de imágenes e identificación de nanoagregados. Además, hemos extraído atributos existentes en la literatura para la caracterización del *carbon black* y para en análisis de regiones en general. Adicionalmente, hemos diseñado nuevas características. Todas ellas han sido validadas mediante cuatro métodos de selección de atributos, dos univariables y dos multivariables. Así, con el *carbon black* como caso de uso, hemos demostrado la

validez de nuestras características propuestas, algunas de las cuales han superado a todas las existentes en la literatura, con el conjunto TEM y el método *embebido* de selección de atributos a partir del clasificador *Random Forest*.

2. **El primer sistema para la identificación de las categorías morfológicas de agregados de *carbon black* mediante técnicas de Inteligencia Artificial.** Hemos propuesto la aplicación de diferentes algoritmos de aprendizaje automático para la clasificación en las cuatro categorías definidas por Herd [HMH92]. La evaluación ha sido realizada mediante la precisión de acierto y mediante el AUC (Área por debajo de la curva ROC). Se ha establecido como referencia la precisión de acierto media de 6 expertos respecto a la clasificación según sus 6 votos. Ésta ha sido del 80 %, lo que ha sido superado por nuestra clasificación automática.

3. **Un método para la reconstrucción 3D de agregados de *carbon black*.** Hemos propuesto un método de reconstrucción 3D a partir de sólo dos imágenes ortogonales de microscopio TEM. Utilizando la información de densidad a partir de la intensidad de la escala de grises en cada punto del agregado y mediante un algoritmo genético obtenemos resultados comparables a las reconstrucciones tomográficas, las cuales emplean casi un centenar de imágenes. La solución se valida mediante una función de *fitness* o ajuste que indica la similitud del agregado reconstruido desde el punto de vista de las dos imágenes de partida. Adicionalmente, se valida con una tercera imagen a 45º de estas dos imágenes, que no ha sido usada para la reconstrucción. Este enfoque ha resultado efectivo y mucho más eficiente que las técnicas actuales.

4. **Una técnica para la extracción de características de un modelo 3D.** A partir de las posiciones y tamaño de las partículas obtenidas del algoritmo genético hemos extraído información de utilidad. Para calcular el volumen del agregado hemos transformado la representación basada en esferas en una representación volumétrica mediante una red de *voxelización*. Además, para calcular la superficie hemos transformado la representación volumétrica en una de *isosuperficie* gracias a nuestra implementación mejorada del algoritmo *Marching Cubes* [LC87]. Estas características y el número de partículas han sido validadas mediante la reconstrucción tomográfica y una simulación basada en Montecarlo.

Gracias a estas contribuciones hemos cumplido los objetivos específicos establecidos en el capítulo 1:

- *Desarrollar y evaluar un caracterizador de agregados de negro de humo por medio de técnicas de selección de atributos.*
- *Desarrollar y evaluar un categorizador de agregados de negro de humo por medio de técnicas de machine learning.*
- *Desarrollar y evaluar un reconstructor 3D de agregados que permita estudiar su estructura real y obtener características adicionales.*

Igualmente, hemos cumplido los objetivos operacionales que surgen de éstos:

- *Diseñar e implementar un sistema que realice un tratamiento sobre una imagen de microscopio electrónico que permita extraer información relevante de ella.*
- *Diseñar e implementar un sistema que extraiga información de una imagen que contenga un agregado segmentado binarizado.*
- *Diseñar e implementar un sistema que asista en el etiquetado manual de agregados.*
- *Diseñar e implementar un sistema que clasifique agregados por medio de algoritmos de machine learning usando las características extraídas de los agregados.*
- *Validar el categorizador desarrollado incluyendo un estudio de la relevancia de las características morfológicas.*
- *Diseñar e implementar un sistema que sea capaz de reconstruir tridimensionalmente un agregado por medio de sólo dos imágenes TEM, así como validarlo con reconstrucciones tomográficas.*
- *Diseñar e implementar un sistema que obtenga nuevas características de la reconstrucción tridimensional.*
- *Diseñar e implementar un sistema que permita realizar controles de calidad de manera automatizada por medio de la categorización de agregados.*

- *Diseñar e implementar un sistema que permita analizar las características de un material para el control de la producción.*

Tras haber cumplido estos objetivos operacionales, consideramos superado el objetivo principal: *"Desarrollar y evaluar un caracterizador y categorizador de agregados de negro de humo que permita realizar un control de calidad de nanomateriales y asista en los procesos de producción de materiales con nuevas propiedades."* y por tanto la hipótesis inicial planteada. Gracias a las características existentes, y a las que hemos aportado, somos capaces de clasificar los agregados de *carbon black* de manera automática, un proceso de gran utilidad para las tareas de control de calidad y la investigación para la producción de nuevos materiales. Además, hemos presentado un método innovador de reconstrucción tridimensional con sólo dos imágenes, a partir del cual podemos conocer la morfología tridimensional precisa y extraer características que si no serían estimadas.

5.3 Aplicaciones de la investigación

Las aplicaciones de esta investigación se centran en el estudio de la morfología de los nanomateriales, en concreto del *carbon black*, aunque también incluyen el análisis de objetos a escala real.

Así, la aplicación más directa es el estudio de la distribución frecuencial de las morfologías de un nanomaterial dado, con el objetivo de llevar a cabo un control de su calidad. Cabe recordar, que cada forma proporciona unas propiedades diferentes al material con el que es mezclado, por lo que, saber las proporciones de cada tipo, es de gran utilidad. Aun así, el resto de atributos pueden utilizarse también para comprobar que un proveedor sigue proporcionando un material con las mismas características con las que se está acostumbrado a trabajar.

Por otro lado, en las fases de creación de nuevos materiales, es de gran utilidad tanto la clase morfológica como el conjunto de características extraídas, para analizar su relación con las propiedades mecánicas de cada material creado. Asimismo, gracias a la reconstrucción 3D será posible el estudio detallado de la morfología.

Además, mediante la metodología propuesta de selección de atributos, se facilitará la evaluación de nuevos atributos que se diseñen en un futuro para mejorar la caracterización. Igualmente, permitirá la reducción del número de

atributos con el objetivo de agilizar los procesos de caracterización y categorización. Como ya se ha explicado anteriormente, habrá que conocer el volumen de trabajo y los requisitos temporales para decidir si es necesario reducir la cantidad de atributos y encontrar el equilibrio entre efectividad y eficiencia.

El caso de uso empleado ha sido el *carbon black*, sin embargo, la metodología empleada permite el estudio de otros materiales. Así, se podrían definir nuevas características y evaluarse mediante los métodos de selección de atributos expuestos, aunque esto no sería necesario. El trabajo imprescindible sería el etiquetado del material en las morfologías deseadas.

Resumiendo, a pesar de que las aplicaciones directas están enfocadas al *carbon black*, con un esfuerzo relativamente pequeño, podrán extenderse a otros materiales, lo que se expondrá en la sección 5.5 sobre el trabajo futuro.

5.4 Limitaciones de la solución

Las diferentes contribuciones tienen actualmente una serie de limitaciones, que inhabilitan su utilización en algunos casos concretos. A continuación, se enuncian los problemas previstos por las restricciones actuales y posteriormente, en la sección 5.5 se presentan las soluciones previstas.

Primero, las condiciones de captura de las imágenes influyen en gran medida en los resultados y las imágenes de microscopio ya de por sí contienen mucho ruido. Por esto, si no son tomadas por personal experto la calidad de la posterior *binarización* podría no ser aceptable. Hay que tener en cuenta que una *binarización* incorrecta puede provocar que se descarten del estudio parte de los agregados, sesgando los resultados. Así, podrían descartarse los más pequeños, que son los que tienen más probabilidades de ser esferoidales. Por otro lado, en caso de no haber dispersado efectivamente el material podría estar aglomerado, eso es, que los agregados estén unidos y parezca que hay agregados más grandes de lo que son en realidad.

El algoritmo genético de reconstrucción 3D presentado, a pesar de superar enormemente los tiempos requeridos mediante las tomografías, sigue requiriendo de cierta interacción humana y su captura sigue siendo más costosa que el realizar un análisis bidimensional de los agregados. Este coste, es, tanto de trabajo manual, como de procesamiento computacional. Será necesario valorar en cada caso la información adicional que proporciona.

5.5 Trabajo futuro

Hemos identificado numerosas líneas de trabajo futuras, unas tratan de mejorar la herramienta y otras amplían sus aplicaciones posibles. A continuación, se enuncian las más relevantes.

Por un lado, para mejorar la *binarización* y posterior segmentación, se planea el estudio de la *Cohomología de Čech* [BM11b], técnica que consiste en la elección de una serie de puntos del agregado, asignando un número de éstos más elevado en las zonas más densas y posteriormente expandiéndolos hasta un radio determinado. Además, para mejorar el proceso de dilatado y erosión que mejora la región binarizada, vamos a utilizar discos de tamaño proporcional a los aumentos de la imagen. Otra alternativa planificada es redimensionar las imágenes a un tamaño de píxel uniforme con anterioridad a todo procesamiento. Esto es especialmente importante para la esqueletonización, ya que el tamaño de la región influye enormemente en el grado de ramificación del esqueleto.

Por otro lado, queremos crear un sistema que aplique un tratamiento diferente y con unos parámetros óptimos en función de las características de la imagen. De hecho, hemos probado con éxito diferentes filtros, pero algunos tienen el inconveniente de que también empeoran el procesado de ciertas imágenes. Entre ellos, destacan el filtro de difusión anisotrópica [PM90], con diferentes condiciones de parada [BSMH98, FH01, SBH03, TB10], que es un método para eliminar el ruido que a su vez mantiene los bordes, y el de *wiener* [Wei49], un filtro adaptativo también para la eliminación del ruido, del que existe una mejora para la detección de partículas en imágenes de microscopio [SG11]. Por esto, queremos entrenar un modelo que realice un análisis previo de cada imagen y decida la configuración óptima para *binarizarla*.

Como se ha comentado, el proceso de esqueletización requiere todavía de una optimización que permita su aprovechamiento para discernir entre las cuatro categorías morfológicas. A pesar de haber realizado ajustes para evitar obtener un número excesivo de ramas, su relevancia sigue siendo muy baja.

De la misma forma, se ha detectado el problema de que las clases no estén balanceadas y se planean varios enfoques para superarlo. La solución ideal sería tener un conjunto de imágenes muy grande y con las clases balanceadas. Precisamente, ya se ha creado un conjunto de datos artificial y se ha observado como los errores de clasificación son más uniformes en las cuatro categorías y tanto el *accuracy* como el AUC mejoran.

En concreto, queremos aplicar técnicas de *Active Learning* o aprendiza-

je activo con las que se va proporcionando un subconjunto de las muestras al modelo iterativamente [EHG07]. Para conjuntos con un número elevado de instancias este algoritmo tiene un alto coste computacional, pero existen alternativas para alcanzar resultados parecidos a un coste mucho más reducido [EHG07]. Además, gracias al *Active Learning* también es posible crear un sistema que decida qué imágenes son más relevantes para su etiquetado y minimizar de esta forma este trabajo manual [TC01].

Adicionalmente, creemos que mejoraríamos los resultados de la categorización de agregados con un *ensemble* de clasificadores [Die00]. Esto es, un conjunto de clasificadores que votan para decidir la categoría a elegir.

En cuanto a la reconstrucción 3D, queremos llevar a cabo validaciones adicionales con agregados de *carbon black* más complejos y con múltiples agregados simultáneamente. Ya hemos trabajado en la reconstrucción de varios agregados, sin embargo, no se han obtenido resultados satisfactorios todavía. Se han establecido diferentes penalizaciones para la función de *fitness*, pero es necesario establecer restricciones adicionales para conseguir que el algoritmo converja hacia una solución correcta. De esta forma se permitirá un estudio mucho más eficiente y con menos trabajo manual de captura de imágenes. Asimismo, se definirán nuevas características para extraer información adicional de la reconstrucción, que en un futuro podrían alimentar a un algoritmo de aprendizaje automático y mejorar la calidad obtenida mediante el análisis bidimensional.

Además, también tenemos intención de adaptar el algoritmo para la reconstrucción de otros materiales, reduciendo el tamaño de partícula para reconstruir objetos que no estén formados por partículas esféricas. Esto lo haría más costoso computacionalmente por lo que sería conveniente optimizar el algoritmo.

Entre las optimizaciones planificadas se encuentra la codificación para GPGPU de las partes críticas. Así, gracias al aprovechamiento del procesador de la tarjeta gráfica será posible paralelizar el trabajo, lo que es totalmente viable por la naturaleza de los algoritmos genéticos.

Por último, el concepto utilizado por el algoritmo genético para la reconstrucción tridimensional de nanoagregados, podría ser utilizado para aprovechar la información de intensidad en diferentes tipos de imágenes, como pueden ser las imágenes térmicas o las obtenidas mediante rayos-X.

5.6 Consideraciones finales

Actualmente, están más extendidos los métodos indirectos de caracterización de nanomateriales. Éstos incluyen la absorción de aceite y la adsorción de nitrógeno y son más rápidos y permiten el estudio de conjuntos más grandes. En cambio, los métodos directos, que consisten en la captura del material mediante imágenes de microscopio para su posterior procesamiento, han sido dejados un poco de lado especialmente en los entornos industriales.

Por esto, en la presente tesis doctoral, hemos trabajado para automatizar los métodos directos y hacerlos más robustos frente a las imágenes de microscopio, que por su naturaleza contienen mucho ruido. Así, con el progreso realizado hemos alcanzado la precisión de acierto de esta tarea realizada de manera manual.

Hay que procurar sistematizar los procesos de captura para obtener unas imágenes con condiciones similares. Sin embargo, el sistema tiene que ser muy flexible, ya que, cada material ha podido tener un tratamiento previo diferente, como puede ser disolución en cloroformo, que deje residuos no deseados que dificulten su tratamiento. Además, cada microscopio puede aportar sus particularidades e incluso el estado del filamento de la pistola que lanza el haz de electrones influye en las imágenes obtenidas.

Además de las mejoras en la *binarización* y caracterización, si se automatiza completamente el proceso de captura, mediante un microscopio motorizado con capacidad de tomar muestras de manera autónoma, los métodos directos podrían desbancar a los indirectos, proporcionando información mucho más completa del material analizado.

Por último, el método de reconstrucción 3D presentado permitirá realizar reconstrucciones de *carbon black* y otros materiales a un coste y con una duración infinitamente menores que con los métodos actuales.

> ≪*El cambio es ley de vida. Cualquiera que sólo mire al pasado o al presente, se perderá el futuro.*≫
>
> John Fitzgerald Kennedy (1917-1963)

"Future is amazing, but it needs to be made."

Elisa Carrió (1956-present)

6

Conclusions

Once we have exposed the research process followed, the solutions provided and their evaluation, it is the time to revisit the hypothesis and initial objectives to check their degree of achievement. In addition, in this chapter, we will present the limitations of the proposed system and the identified future work in the three topics of this research about carbon black: characterization, categorization and 3D reconstruction.

In this way, the chapter is organized as follows. Section 5.1 presents the different validations of the conducted experiments. Next, section 5.2 summarizes the obtained results, and compares them with the initial hypothesis and the objectives that come from it. Later, section 5.3 shows the identified applications for which our contributions will mean a competitive advantage. Moreover, section 5.4 reveals the limitations found. Furthermore, section 5.5 explains the solution to those limitations, and defines the future lines of research. Lastly, section 5.6 concludes this thesis and provides the final considerations of the author.

6.1 Synthesis

In this work, we have conducted several empirical validations of the experiments. The conclusions reached are presented below.

First, we validate the new features we have designed to extract information from aggregate microscopy images. We employ four feature selection algorithms, two univariate and two multivariate. Moreover, we have used three datasets, one of SEM (Scanning Electron Microscope) images, one of TEM (Transmission Electron Microscopy) images and one of artificial images made so we could have a class balanced dataset. SEM and TEM datasets are not balanced, which can be a problem with the traditional machine learning algorithms [EHG07]. Nevertheless, it is worth pointing out that we consider the TEM dataset to be more relevant, for its size and its diversity in the image capturing conditions. In addition, it is the most common technique employed for analyzing nanometric specimens [RS06].

To assess the relation between the features and the morphology, we have manually tagged aggregates into the four categories defined by Herd [HMH92]. This task is not trivial, because differences between classes are not clearly defined, and it is a bit subjective. Besides, we have used two univariate methods: Chi-squared and Information Gain Ratio. According to these methods, the best feature is *relaciónÁreaÁreaConvexa*, which is commonly used in the literature [PI97, Rus07]. According to Chi-squared, the next feature is *áreaÁreaBoundingBox*, previously used in the literature in image treatment tasks, but not with carbon black.

In addition, the results have proved how good our features are, some of them overcoming the ones present in the literature. This is especially true according to the two multivariate methods: *ReliefF* and an embedded filter with the classifier Random Forest.

In particular, *porcentajeSegmentos* and *porcentajeCortesSegmentos* are the best parameters, overcoming all of the literature ones with the TEM dataset and the feature selection algorithm based on Random Forest, the most relevant method for us, because it has provided the best classification results.

In general, *mediaÁreaÁreaConvexa* and *cambioÁreaSuavizada*, also designed by us, have proven relevant in most of the cases. Then, with still a good result, are *desviaciónÁreaÁreaConvexa, mediaÁreaÁreaTriángulo* and *desviaciónÁreaÁreaTriángulo*. Lastly, six of our features have yield bad results. These are *sectoresVacíos, cambiosDerivada* and the ones related with the skeletonization: *desviaciónRamas, medianaRamas, mediaRecortadaRamas* and *mediaWindsorizadaRamas*.

These last four features we have designed depend on a skeletonization process that needs optimizing. In the literature it has been used with success

[MG99] and it is clear that we can extract morphology information with this type of features.

Regarding the processing times, the feature selection algorithm based on Random Forest, exceeds the rest by far, especially for the one with 1,000 trees. Nevertheless, it is still acceptable, because it is not normally an urgent task. It can be used for reducing the dimensionality of a dataset with the aim of creating a classification model. However, it is possible that in some cases we may prefer to slightly reduce the quality in exchange for a notable decrease of time. This can be true with big datasets or in the case where the feature selection process is chained with the generation of a classification model that is frequently updated.

Meanwhile, we have performed an automatic categorization of carbon black aggregates by morphology type with the three datasets mentioned for the characterization process. We have used four base classifiers with different kernels and configurations, making a total of 17 algorithms. The results evaluation has been carried out by means of the accuracy and AUC (Area Under the ROC Curve). The AUC has the advantage of not being affected by the class imbalance problem, in contrast to other metrics like the accuracy [Faw06].

Since in the literature there are no measures of the quality of the classifications and it is a task with a high degree of subjectivity, we have conducted independent ratings of 6 experts and aggregates have been classified by vote. Subsequently, we have compared each individual classification with the one made by vote and the average score was 80 % accuracy. This value was considered as the threshold necessary to categorize automatically.

On the one hand, for the SEM dataset, the best algorithm in terms of accuracy has been the Random Forest of a 100 trees, with a 79.5 %. On the contrary, the one with a 1,000 trees has the best AUC, 0.931. The worst machine learning algorithm is the KNN of 1 neighbour, followed by Naïve Bayes classifier. No classifier overcomes the four decision trees, and the SVM (Support Vector Machine) with the Pearson VII Universal Kernel is the best in terms of accuracy of the rest of classifiers. According to AUC, KNNs of 20 and 25 neighbours follow the three configurations of Random Forest tested.

On the other hand, for the TEM dataset, we have noticed an improvement in all the algorithms, with the exception of J48 decision tree. Specifically, the three Random Forests and the three SVMs yielded an accuracy over 80 %. One of the main reasons is the size of the dataset. If we reduce the dataset to the SEM size, the algorithms effectiveness is reduced. The Random Forest of

1,000 trees is again the best with 81.69 % accuracy and 0.94 of AUC. The 100 trees version and the SVM with polynomial kernel follow it closely, although, with the SEM dataset the SVM results were not especially good. Additionally, the three Random Forest and all KNNs except from the one with 1 neighbour, surpass the 0.9 of AUC.

In contrast, for the artificial dataset, the Pearson VII Universal Kernel stands out with 85.07 % accuracy. According to AUC, the Random Forest is the best algorithm again. By means of the confusion matrixes, we see how with this dataset, spheroidal aggregates are correctly classified, with only 5% of them incorrectly identified. In this way, having balanced classes helps the categorization of this class, which is a minority in our other two datasets. Moreover, thanks to confusion matrixes we have appreciated the difficulty in discerning between the linear aggregates and the ellipsoidal or branched ones.

With regards to the processing times, KNNs and Bayes classifiers are faster and the Random Forest of 1,000 trees needs far more time than the rest. Concretely, it takes 200 times more to generate the model than the Naïve Bayes algorithm, and 40 times more to classify an instance. Nevertheless, the required times are not excessive and should not imply any problem. If that were the case, the number of trees or the least relevant features should be reduced.

The last contribution, the 3D reconstruction by means of a genetic algorithm has been validated with the classic tomography reconstruction and with a Monte Carlo simulation. The designed method has obtained satisfactory results in far less time, far less manual work and with only two orthogonal images.

In only one hour, we have obtained good results, and with more time it continues improving. In addition, we have designed a validation method with an additional image that has not been used for the aggregate reconstruction.

6.2 **Summary of results**

This section summarizes the results obtained during this investigation to validate the hypothesis, which states:

It is possible to develop an automated system for the quality control of nanomaterials composed of carbon black and to improve its production process, with microscope image processing, characterization and subsequent application of Machine Learning techniques.

The contributions in the area of carbon black aggregates characterization, categorization and 3D reconstruction are summarized below. Their validations have been compiled in section 6.1. Thanks to them we enable automated analysis of their morphology, a very useful task for both quality control and for the manufacture of new materials with improved properties.

1. **A method for extracting features from nanoaggregates.** We have proposed a technique for image treatment and identification of nanoaggregates. In addition, we have extracted attributes from the literature about carbon black characterization and about analyzing regions in general. Additionally, we have designed new features. All have been validated by four feature selection methods, two univariate and two multivariate. So, with the carbon black as a use case, we have demonstrated the validity of our proposed features, some of which have outperformed all existing in the literature, with the TEM dataset and the embedded filter with the classifier Random Forest.

2. **The first system for the identification of morphological categories of carbon black aggregates through Artificial Intelligence techniques.** We have proposed applying different machine learning algorithms to classify carbon black aggregates into the four categories defined by Herd [HMH92]. The evaluation was carried out with the accuracy and AUC (Area Under the ROC Curve). We have compared the accuracy with the average of 6 experts with regard of the classification according to its 6 votes. This was 80 %, which has been overtaken by our automatic classification algorithm.

   Reference is established wisdom accuracy average of 6 experts regarding the classification according to its 6 votes. It has been 80 %, which has been overtaken by our automatic classification.

3. **A 3D reconstruction method for carbon black aggregates.** We have proposed a 3D reconstruction method from only two orthogonal

TEM images. We use the density information from the intensity of the greyscale at each point of the image. Moreover, we have developed a genetic algorithm to obtain comparable results to tomographic reconstructions, which employ almost a hundred images. The solution is validated by a fitness function indicating the similarity of the reconstructed aggregate from the viewpoint of the two initial images. Additionally, it is validated with a third image taken at 45º of these two images, which has not been used for the reconstruction. This approach has been effective and more efficient than current techniques.

4. **A technique for extracting features from a 3D model.** From the position and size of the particles obtained from the genetic algorithm, we have extracted useful information. To calculate the volume of the aggregate we have transformed the spheres representation into a volumetric representation through a *voxelization* network. Furthermore, to calculate the surface we have transformed the volumetric representation into an *isosurface* representation, enhanced by our implementation of the *Marching Cubes* algorithm [LC87]. These features and the number of particles have been validated by a tomographic reconstruction and a Monte Carlo based simulation.

Thanks to these contributions we have met the specific objectives set out in chapter 1:

- *Developing and evaluating a characterization system of carbon black aggregates using feature selection techniques.*
- *Developing and evaluating a categorization system of carbon black aggregates through machine learning techniques.*
- *Developing and evaluating a three-dimensional reconstruction system that enables studying its actual structure and obtaining additional features.*

Similarly, we have met the operational objectives arising from these:

- *Designing and implementing a system to perform a treatment on an electron microscope images that enables extracting relevant information from it.*

- *Designing and implementing a system to extract information from an image containing a segmented binarized aggregate.*
- *Designing and implementing a system to assist in the manual labelling of aggregates.*
- *Designing and implementing a system that classifies aggregates with machine learning algorithms using features extracted from the aggregates.*
- *Validating the categorizer developed including a study of the relevance of the morphological features.*
- *Designing and implementing a system that is able to reconstruct an aggregate three-dimensionally using only two TEM images and validating with tomographic reconstructions.*
- *Designing and implementing a system to get the new features of three-dimensional reconstruction.*
- *Designing and implementing a system to carry out exhaustive quality controls through automated aggregate categorization.*
- *Designing and implementing a system to analyze the characteristics of a material for production control.*

Having completed these operational objectives, we consider the main objective accomplished: *"Developing and evaluating a carbon black aggregates characterizer and categorizer that enables quality control of nanomaterials and assists in the production processes of new materials properties."* and therefore the initial hypothesis posed. Thanks to existing features, as well as the ones we have contributed, we are able to classify the carbon black aggregates automatically. This process is extremely useful for quality control tasks and for the research on new materials production. In addition, we have presented an innovative method of three-dimensional reconstruction with only two images. With it, we can determine the precise three-dimensional morphology, and extract features, which otherwise would be estimated.

6.3 Research applications

The applications of this research are related to the study of nanomaterials morphology, especially carbon black, but also include the analysis of objects in real scale.

Thus, the most direct application is the study of the frequency distribution of the morphologies of a given nanomaterial, with the aim of conducting a quality control. Each morphology provides different properties to the material that it is mixed with, so knowing the proportions of each type is very useful. Still, the other attributes can also be used to verify that a supplier continues to provide a material with the same characteristics with which they are used to work.

On the other hand, in the phases of creating new materials, it is useful to know the morphological class as well as the set of extracted features. These can be related to the mechanical properties of each material created. Also, thanks to the 3D reconstruction it is possible to conduct a detailed study of the morphology.

Furthermore, using the proposed methodology for selecting attributes will facilitate the evaluation of features designed in the future to improve the characterization. This will enable the reduction of the number of attributes in order to speed up the process of characterization and categorization. As explained before, we will need to know the workload and time requirements to decide whether to reduce the number of attributes and find the balance between effectiveness and efficiency.

The carbon black has been the use case; however, the proposed methodology enables the study of other materials. Thus, new features may be defined and evaluated by the attribute selection methods explained, although this would not be necessary. The inevitable work is the labelling of the material in the desired morphologies.

In summary, although direct applications are focused on carbon black, with relatively little effort, they can be extended to other materials, which will be discussed in section 6.5 on future work.

6.4 Limitations of the model

The presented contributions have currently a number of limitations. So in same cases they are not useful. Below are the problems foreseen by the current restrictions and later, in section 6.5 are the solutions provided.

First, the image capturing conditions greatly influence on the results and microscope images already contain a lot of noise. Because of this, if they are not taken by microscopy experts, later, the *binarization* quality might not be acceptable. Keep in mind that an incorrect *binarization* can cause to

discard some aggregates, thus skewing the results. So we could unintentionally discard the smallest ones, which are more likely to be spheroidal. On the other hand, in case the material has not been effectively dispersed, it could form agglomerates, that is, that the aggregates are connected and they seem to be larger than they actually are.

The presented genetic 3D reconstruction algorithm has greatly exceeded the time required by tomography. Still, it requires some human interaction and it is still has more cost than the two-dimensional aggregate analysis. This cost is both manual work and computer processing. It will be necessary to assess in each case the additional information it provides.

6.5 Future Work

We have identified several future research lines. Some of them try to improve the system and others expand its potential applications. The most relevant are presented below.

On the one hand, to improve the *binarization* and subsequent segmentation, we have planned to study the *Čech cohomology* [BM11b]. This technique consists in choosing a number of points from an aggregate, assigning more in denser areas, and later, expanding them to a certain radius. Moreover, to improve the process of dilation and erosion of the binarized region, we are going to use discs of size proportional to the image magnification. Another alternative is to resize the image prior to any processing. This is especially important for the skeletonization because the size of the region determines enormously the degree of ramification.

Furthermore, we want to create a system that applies a different treatment with optimum parameters according to the characteristics of the image. In fact, we have successfully tested different filters, but some of them worsen the processing of certain images. Among them, stands out the anisotropic diffusion filter [PM90], with various stop conditions [BSMH98, FH01, SBH03, TB10]. This method eliminates the noise while preserving the edges. The *wiener* filter [Wei49] is an adaptive filter also for eliminating noise, and there is an improvement for particle detection in microscope images [SG11]. So we want to train a model, to perform a preliminary analysis of each image and decide the optimal settings for its *binarization*.

As aforementioned, the skeletonization process still requires an optimization that enables its use to distinguish between the four morphological cate-

gories. Despite having made adjustments to avoid getting too many branches, its relevance is still very low.

Similarly, the class imbalance problem has been detected and we have planned several approaches to overcome it. The ideal solution would be to have a set of very large images with balanced classes. Precisely, we have created an artificial data set and have observed how classification errors distribute more uniformly in the four categories, and the accuracy and AUC improve.

Specifically, we want to apply *Active Learning* techniques, which provide the model with a subset of samples iteratively [EHG07]. For datasets with a large number of instances, this algorithm has a high computational cost, but there are alternatives to achieve similar results at a much lower cost [EHG07]. And also, thanks to the *Active Learning*, it is also possible to create a system that decides which images are most relevant for labelling and minimize this manual task [TC01].

Additionally, we believe that the aggregate categorization results would improve with an *ensemble* of classifiers [Die00]. That is, a set of classifiers that vote to decide the category you choose.

For 3D reconstruction, we want to perform additional validations with more complex carbon black aggregates and with multiple aggregates simultaneously. We have worked on the reconstruction of several aggregates; however, no satisfactory results have been obtained yet. We have established different penalties for the fitness function, but it is still necessary to establish additional restrictions to ensure that the algorithm converges to a correct solution. This will enable a much more efficient research and with less manual imaging labour. In addition, new features will be defined to extract information of the reconstruction, which in the future could feed a machine learning algorithm and improve the quality obtained by bidimensional analysis.

Furthermore, we intend to adjust the algorithm for the reconstruction of other materials, reducing the particle size to reconstruct objects that are not formed by spherical particles. This would make it more computationally expensive and it would be desirable to optimize the algorithm.

One of the planned performance enhancements is GPGPU coding of critical parts. So, thanks to the exploitation of graphic card processor, we will be able to parallelize the work, which is entirely feasible because of the nature of genetic algorithms.

Finally, the concept used by the genetic algorithm for three-dimensional reconstruction of nanoaggregates might be used to get current information

on different types of images, such as thermal imaging or those obtained by X-rays.

6.6 Final considerations

Currently, nanomaterials indirect characterization methods are more widespread. These include oil absorption and nitrogen adsorption, and are quicker and enable the study of larger sets. In contrast, direct methods, consisting in the capture of the material by microscopic images for further processing, are left aside especially in industrial environments.

Therefore, in this dissertation, we have worked to automate direct methods and make them more robust to the microscope images, which because of their nature contain a lot of noise. Thus, with the progress already made we have reached the accuracy obtained doing this task manually.

Efforts should be made to systematize the capture processes to obtain images with similar conditions. However, the system has to be very flexible, since each material may have had a different pre-treatment, as the dissolution in chloroform, which may leave unwanted residues which hinder their treatment. Moreover, each microscope can produce peculiar artefacts, and even the state of the filament of the gun that launches the electron beam, influences the obtained images.

In addition to improvements in the *binarization* and characterization, if the process could be automated entirely, including the use of a motorized microscope capable of independent sampling, direct methods could replace indirect ones, providing more complete information of the analyzed material.

Finally, the 3D reconstruction method presented will allow reconstructions of carbon black and other materials at a much lower money and time cost than with current methods.

> "*Change is the law of life. And those who look only to the past or present are certain to miss the future.*"
>
> John Fitzgerald Kennedy (1917-1963)

Publicaciones

En el transcurso de la presente investigación se han realizado varias publicaciones que se enumeran a continuación:

Revistas

1. Juan López-de-Uralde, Iraide Ruiz, Igor Santos, Agustín Zubillaga, Pablo G. Bringas, Ana Okariz y Teresa Guraya. *Automatic Morphological Categorisation of Carbon Black Nano-aggregates*. En Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, volumen 6262, páginas 185-193, 2010. (Del congreso DEXA).

Artículos en congresos

2. Juan López-de-Uralde, Iraide Ruiz, Igor Santos, Agustín Zubillaga, Pablo G. Bringas, Ana Okariz y Teresa Guraya. *Automatic Morphological Categorisation of Carbon Black Nano-aggregates*. En Proceedings of the 21^{st} International Conference on Database and Expert Systems Applications (DEXA), volumen 2, páginas 185-193. Bilbao, 30 de agosto al 3 de septiembre de 2010. Rank B en CORE-ERA.

3. Juan López-de-Uralde, Teresa Guraya, Ana Okariz, Enkarni Gómez, Agustín Zubillaga e Igor Santos. *Automated Image Analysis and Artificial Intelligence applied to microscopic images: a new methodology for the morphological characterization of nano-particles*. En Proceedings of the Modelling of Elastomeric Materials and Products. Londres, 14 de octubre de 2010.

4. Juan López-de-Uralde, Mikel Salazar, Aitor Santamaría, Agustín Zubillaga, Pablo G. Bringas, Teresa Guraya, Ana Okariz, Encarni Gómez

y Zineb Saghi. *Three-dimensional carbon black aggregate reconstruction from two orthogonal TEM images*. En Proceedings of the 7^{th} European Conference on Constitutive Models for Rubber (ECCMR). Dublin, Ireland, páginas 433-438, 20 al 23 de septiembre de 2011.

5. Julen Ibarretxe, Maider Iturrondobeitia, Floren Garrido, Juan López-de-Uralde, Zineb Saghi. *Obtaining 3D structural models from TEM-tomography reconstructions*. En Proceedings of the 2^{nd} Joint Congress of the Portuguese and Spanish Microscopy Societies. Aveiro, Portugal, 18 al 21 de octubre de 2011.

6. Maider Iturrondobeitia, Juan López-de-Uralde, Julen Ibarretxe y Floren Garrido. *Innovative microscopy based techniques for 2D and 3D structural characterization of polymer composites*. En Proceedings of the 11^{th} European Symposium on Polymer Blends. San Sebastián, 25 al 28 de marzo de 2012.

7. Maider Iturrondobeitia, Teresa Guraya, Julen Ibarretxe,Ana María Zaldua y Juan López-de-Uralde. *Relevance of processing parameters and structure of layered silicate bionanocomposites on their final application properties*. En Proceedings of the 15^{th} European Conference on Composite Materials (ECCM15). Venice, Italy, 24 al 28 de junio de 2012. Congreso A.

Charlas invitadas

8. *Categorization of Carbon Black Aggregates from New Characteristics and Machine Learning Algorithms*. Abstract and biography in proceedings of the BIT's 1^{st} Annual World Congress of Nano S&T at the World EXPO Center. Dalian, China, volumen 2, página 751, 23 al 26 de octubre de 2011.

Talleres

9. *Digital images processing for better quality in reconstruction*. Introduction to tomography technique in TEM, workshop between invited researchers from Cambridge University, Cádiz University, CIC bioGUNE (Center for Cooperative Research in Biosciences), University of the Basque Country and University of Deusto. San Sebastián, 6 al 7 de octubre de 2009.

Bibliografía

[AA90] K. M. Abraham y M. Alamgir. Li+-Conductive solid polymer electrolytes with liquid like conductivity. *Journal of The Electrochemical Society*, 137(5):1657–1658, 1990.

[Abu89] A. S. Abutaleb. Automatic thresholding of gray-level pictures using two-dimensional entropy. *Computer Vision, Graphics and Image Processing*, 47(1):22–32, 1989.

[AGD75] A. A. Abrikosov, L. P. Gor'kov, y I. E. Dzyaloshinski. *Methods of quantum field theory in statistical physics*. Dover Pubns, 1975.

[Ali12] Alicona. MeX turns any SEM into a 3D measurement device. Disponible en: `http://www.alicona.com/home/products/mex.html`, 2012.

[Ama03] A.L. Amaral. *Image analysis in biotechnological processes: applications to wastewater treatment*. Proyecto Fin de Carrera, Universidade do Minho, Portugal, 2003.

[Ame02] American Society for Testing and Materials. *ASTM D3849-02 - Standard Test Method for Carbon Black - Morphological Characterization of Carbon Black Using Electron Microscopy*, 2002. Testing method.

[Ame07] American Society for Testing and Materials. *ASTM D3849-07 - Standard Test Method for Carbon Black - Morphological Characterization of Carbon Black Using Electron Microscopy*, 2007. Testing method.

BIBLIOGRAFÍA

[Ame10] American Society for Testing and Materials. *ASTM D6556 - 10 Standard Test Method for Carbon Black-Total and External Surface Area by Nitrogen Adsorption*, 2010.

[Ame11] American Society for Testing and Materials. *ASTM D2414 - 11 Standard Test Method for Carbon Black—Oil Absorption Number (OAN)*, 2011.

[AMG12] I. Alić, J. Muntermann, y R. W. Gregory. State of the Art of Financial Decision Support Systems based on Problem, Requirement, Component and Evaluation Categories. En *Proceedings of the 25th Bled eConference eDependability: Reliable and Trustworthy eStructures, eProcesses, eOperations and eServices for the Future*. 2012.

[AR05] T. Acharya y A. K. Ray. *Image processing: principles and applications*. Wiley-Interscience, 2005.

[Ass06a] International Carbon Black Association. Carbon Black Uses. Disponible en: `http://www.carbon-black.org/uses.html`, 2006.

[Ass06b] International Carbon Black Association. Health & Hygiene. Disponible en: `http://www.carbon-black.org/health.html`, 2006.

[Ass06c] International Carbon Black Association. What is Carbon Black? Disponible en: `http://carbon-black.org/what_is.html`, 2006.

[AW99] S. Amari y S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[Bar54] N. A. Barricelli. Esempi numerici di processi di evoluzione. *Methodos*, 6(21-22):45–68, 1954.

[Bay63] T. Bayes. An essay towards solving a problem in the doctrine of chances. *MD computing: computers in medical practice*, 8(3):157, 1763.

[BET38] S. Brunauer, P. H. Emmett, y E. Teller. Adsorption of gases in multimolecular layers. *Journal of the American Chemical Society*, 60(2):309–319, 1938.

[BFC99] A. M. Brasil, T. L. Farias, y M. G. Carvalho. A recipe for image characterization of fractal-like aggregates. *Journal of Aerosol Science*, 30(10):1379–1389, 1999.

[BFH+11] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, y D. Scuse. *WEKA Manual for Version 3-7-4*, 2011.

[BFO+84] L. Breiman, J. Friedman, R. Olshen, C. Stone, D. Steinberg, y P. Colla. CART: Classification and regression trees. *Wadsworth: Belmont, CA*, 1984.

[Bis95] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.

[Bis06] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[BK12] D. P. Bhattacharya y S. Koner. English alphabet recognition using chain code and LCS. *Indian Journal of Computer Science and Engineering*, 3(2):329–335, 2012.

[BLNdL95] A. Beghdadi, A. Le-Négrate, y P. V. de Lesegno. Entropic thresholding using a block source model. *Graphical Models and Image Processing*, 57(3):197–205, 1995.

[BM11a] C. Bell y M. Marrapese. Nanotechnology standards and international legal considerations. *Nanotechnology Standards*, páginas 239–255, 2011.

[BM11b] F. Bulnes y J. C. Maya. Cohomology of Moduli Spaces in Differential Operators Classification to the Field Theory. En *Proceedings of the 8^{th} International Conference on Function Spaces, Differential Operators and Nonlinear Analysis*, tomo 1, páginas 1–22. 2011.

[Bor86] G. Borgefors. Distance transformations in digital images. *Computer vision, graphics and image processing*, 34(3):344–371, 1986.

[BR96] M. J. Black y A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996.

[Bre01] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[BRH+06] P. J. A. Borm, D. Robbins, S. Haubold, T. Kuhlbusch, H. Fissan, K. Donaldson, R. Schins, V. Stone, W. Kreyling, J. Lademann, J. Krutmann, D. Warheit, y E. Oberdorster. The potential risks of nanomaterials: a review carried out for ECETOC. *Particle and fibre toxicology*, 3(1):11, 2006.

[Bri89] A. D. Brink. Grey-level thresholding of images using a correlation criterion. *Pattern Recognition Letters*, 9(5):335–341, 1989.

[Bri95] A. D. Brink. Minimum spatial entropy threshold selection. *IEE Proceedings - Vision, Image and Signal Processing*, 142(3):128–132, 1995.

[BSMH98] M. J. Black, G. Sapiro, D. H. Marimont, y D. Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998.

[BTB+05] A. Bogner, G. Thollet, D. Basset, P. H. Jouneau, y C. Gauthier. Wet STEM: A new development in environmental SEM for imaging nano-objects included in a liquid phase. *Ultramicroscopy*, 104(3):290–301, 2005.

[But03] K. Butler. Measuring the oil absorption of carbon black. *Rubber World*, páginas 239–284, 2003.

[Can86] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(6):679–698, 1986.

[Car87] M. J. Carlotto. Histogram analysis using a scale-space approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):121–129, 1987.

[CBHD06] J. Chen, J. Benesty, Y. Huang, y S. Doclo. New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1218–1234, 2006.

[CC99] H. D. Cheng y Y. H. Chen. Fuzzy partition of two-dimensional histogram and its application to thresholding. *Pattern Recognition*, 32(5):825–843, 1999.

[CGH97] E. Castillo, J. M. Gutiérrez, y A. S. Hadi. *Expert systems and probabilistic network models*. Springer Verlag, 1997.

[CH91] G. F. Cooper y E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. En *Proceedings of the 7^{th} conference on Uncertainty in Artificial Intelligence*. 1991.

[CH99] P. A. Ciullo y N. Hewitt. *The rubber formulary*. William Andrew, 1999.

[CL98] J. Cai y Z. Q. Liu. A new thresholding algorithm based on all-pole model. En *Proceedings of the 14^{th} International Conference on Pattern Recognition*, tomo 1, páginas 34–36. IEEE Computer Society, 1998.

[CM88] B. Chanda y D. D. Majumder. A note on the use of the graylevel co-occurrence matrix in threshold selection. *Signal Processing*, 15(2):149–167, 1988.

[CR11] D. Clery y S. Reardon. On the Horizon? European Commission Outlines 80 Billion euros Research Budget. Disponible en: `http://news.sciencemag.org/scienceinsider/2011/11/on-the-horizon-european-commission.html`, 2011.

[CSZ06] O. Chapelle, B. Scholkopf, y A. Zien. *Semi-supervised learning*. Citeseer, 2006.

[CT93] S. C. Cheng y W. H. Tsai. A neural network implementation of the moment-preserving technique and its application to thresholding. *IEEE Transactions on Computers*, 42(4):501–507, 1993.

[DBW93] J. B. Donnet, R. C. Bansal, y M. J. Wang. *Carbon black: science and technology*. CRC Press, 2ª edición, 1993.

[DHS01] R. O. Duda, P. E. Hart, y D. G. Stork. *Pattern classification*. John Wiley, 2001.

[Die00] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, páginas 1–15, 2000.

[dJK02] K. P. de Jong y A. J. Koster. Three-Dimensional Electron Microscopy of Mesoporous Materials-Recent Strides Towards Spatial Imaging at the Nanometer Scale. *ChemPhysChem: A European Journal of Chemical Physics and Physical Chemistry*, 3(9):776–780, 2002.

[Don98] J. B. Donnet. Black and white fillers and tire compound. *Rubber Chemistry and Technology*, 71(3):323–341, 1998.

[dPVU12] Universidad del País Vasco (UPV/EHU). EMERG (EMERG is a Materials Engineering Research Group). Disponible en: `http://www.emerg.es`, 2012.

[DRK68] D. J. De Rosier y A. Klug. Reconstruction of three dimensional structures from electron micrographs. *Nature*, 217(5124):130–134, 1968.

[Ech09] P. Echlin. *Handbook of sample preparation for scanning electron microscopy and x-ray microanalysis*. Springer Verlag, 2009.

[EHG07] S. Ertekin, J. Huang, y C. L. Giles. Active learning for class imbalance problem. En *Proceedings of the 30^{th} annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 823–824. ACM, 2007.

[EM04] E. V. Esquivel y L. E. Murr. A TEM analysis of nanoparticulates in a polar ice core. *Materials Characterization*, 52(1):15–25, 2004.

[Eur84] European Committee for Biological Effects of Carbon Black. *Legislation concerning carbon black*, 1984.

[FA13] A. Frank y A. Asuncion. UCI machine learning repository. Disponible en: `http://archive.ics.uci.edu/ml`, 2013.

[Faw06] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[FC77] W. Frei y C. C. Chen. Fast boundary detection: A generalization and a new algorithm. *IEEE Transactions on Computers*, 100(10):988–998, 1977.

[Fer00] X. Fernandez. Implicit Model-Oriented Optimal Thresholding Using the Komolgorov-Smirnov Similarity Measure. En *Proceedings of the 15th International Conference on Pattern Recognition*, tomo 1, páginas 466–469. IEEE Computer Society, 2000.

[FH01] A. S. Frangakis y R. Hegerl. Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *Journal of structural biology*, 135(3):239–250, 2001.

[FHJ52] E. Fix y J. L. Hodges Jr. Discriminatory Analysis-Nonparametric Discrimination: Small Sample Performance. *California University Of Berkeley*, 1952.

[FKHC$^+$12] B. Fish, A. Khan, N. Hajj Chehade, C. Chien, y G. Pottie. Feature selection based on mutual information for human activity recognition. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 1729–1732. IEEE, 2012.

[FL98] H. Fujiyoshi y A. J. Lipton. Real-time human motion analysis by image skeletonization. En *Proceedings of the IEEE Workshop on Application of Computer Vision*, páginas 15–21. IEEE Computer Society, 1998.

[FNL05] J. Fröhlich, W. Niedermeier, y H. D. Luginsland. The effect of filler–filler and filler–elastomer interaction on rubber reinforcement. *Composites Part A*, 36(4):449–460, 2005.

[FP02] D. A. Forsyth y J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[Fra06] J. Frank. *Electron tomography: methods for three-dimensional visualization of structures in the cell*. Springer, 2006.

[Gar95] S. R. Garner. Weka: The Waikato environment for knowledge analysis. En *Proceedings of the New Zealand Computer Science Research Students Conference*, páginas 57–64. 1995.

[GD88] C. R. Giardina y E. R. Dougherty. *Morphological methods in image and signal processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[GE03] I. Guyon y A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3(7-8):1157–1182, 2003.

[GG08] Z. Ge y Z. Gao. Applications of nanotechnology and nanomaterials in construction. *First International Conference on Construction In Developing Countries (ICCIDC–I)*, páginas 235–240, 2008.

[GGP+97] D. Geiger, M. Goldszmidt, G. Provan, P. Langley, y P. Smyth. Bayesian Network Classifiers. En *Machine Learning*, páginas 131–163. 1997.

[GK08] M. Giersig y G. B. Khomutov. *Nanomaterials for application in medicine and biology*. Springer Verlag, 2008.

[GO91] M. Gerspacher y C. P. O'Farrel. Carbon black is a fractal object: an advanced look at an important filler. *Elastomerics*, 123(4):35–39, 1991.

[GP98] R. Guo y S. M. Pandit. Automatic threshold selection based on histogram modes and a discriminant criterion. *Machine vision and applications*, 10(5):331–338, 1998.

[GPS77] E. Giuliao, O. Paita, y L. Stringa. Electronic character-reading system, 1977. Patente Estados Unidos 4.047.152.

[GWBV02] I. Guyon, J. Weston, S. Barnhill, y V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[Haw04] D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[HGM+00] M. Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, y A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly reviews of biophysics*, 33(4):307–369, 2000.

[HJ02] E. Harris Jr. Information gain versus gain ratio: A study of split method biases. *AMAI*, 2002.

[HMH92] C. R. Herd, G. C. McDonald, y W. M. Hess. Morphology of carbon-black aggregates: fractal versus euclidean geometry. *Rubber Chemistry and Technology*, 65(1):107–129, 1992.

[HMPP00] F. Hofer, C. Mitterbauer, I. Papst, y M. A. Pabst. EFTEM tells us what the Tyrolean Iceman inhaled 5300 years ago. En *EUREM*, tomo 12, páginas 9–14. 2000.

[HMSH93] C. R. Herd, G. C. McDonald, R. E. Smith, y W. M. Hess. The use of skeletonization for the shape classification of carbon-black aggregates. *Rubber Chemistry and Technology*, 66(4):491–509, 1993.

[HMU73] W. M. Hess, G. C. McDonald, y E. Urban. Specific shape characterization of carbon black primary units. *Rubber Chemistry and Technology*, 46(1):204–231, 1973.

[Hol75] J. H. Holland. *Adaptation in natural and artificial systems.* University of Michigan Press, 1975.

[HOS87] L. Halada, G. A. Ososkov, y P. Slavkovsky. Histogram concavity analysis by quasicurvature. *Computers and artificial intelligence*, 6(6):523–533, 1987.

[How07] N. Howe. Contour-Pruned Skeletonization. Disponible en: `http://maven.smith.edu/~nhowe/research/code/`, 2007.

[HS88] L. Hertz y R. W. Schafer. Multilevel thresholding using edge matching. *Computer Vision, Graphics and Image Processing*, 44(3):279–295, 1988.

[HS05] A. Hertzmann y S. M. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 27(8):1254–1264, 2005.

[HT01] D. J. Hand y R. J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[HW95] L. K. Huang y M. J. J. Wang. Image thresholding by minimizing the measures of fuzziness. *Pattern recognition*, 28(1):41–51, 1995.

[IAR96] IARC (International Agency for Research on Cancer). *Printing Processes and Printing Inks, Carbon Black and Some Nitro Compounds*, 1996. Monographs on the Evaluation of Carcinogenic Risks to Humans.

[IIG+11] J. Ibarretxe, M. Iturrondobeitia, F. Garrido, J. López-de-Uralde, y Z. Saghi. Quantitative Microstructural Investigation of Carbon-Black-Filled Rubbers by AFM. *Proceedings of the 2nd Joint Congress of the Portuguese and Spanish Microscopy Societies*, 2011.

[ILES00] I. Inza, P. Larranaga, R. Etxeberria, y B. Sierra. Feature subset selection by Bayesian network-based optimization. *Artificial Intelligence*, 123(1-2):157–184, 2000.

[IMK93] I. F. Imam, R. S. Michalski, y L. Kerschberg. Discovering attribute dependence in databases by integrating symbolic learning and statistical analysis techniques. En *Proceeding of the AAAI-93 Workshop on Knowledge Discovery in Databases, Washington DC*. 1993.

[Ini12] National Nanotechnology Initiative. NNI Budget. Disponible en: `http://www.nano.gov/about-nni/what/funding`, 2012.

[Jäh02] B. Jähne. *Digital Image Processing*. Springer, Berlin, 5ª edición, 2002.

[JBR97] C. V. Jawahar, P. K. Biswas, y A. K. Ray. Investigations on fuzzy thresholding based on fuzzy clustering. *Pattern Recognition*, 30(10):1605–1613, 1997.

[JJ95] D. Janová y J. Jan. Robust surface reconstruction from stereo SEM images. En *Computer Analysis of Images and Patterns*, páginas 900–905. Springer, 1995.

[JJ01] J. Jan y D. Janova. Complex approach to surface reconstruction of microscopic samples from bimodal image stereo data. *Machine Graphics and Vision*, 10(3):261–288, 2001.

[JSK+07] H. Jinnai, Y. Shinbori, T. Kitaoka, K. Akutagawa, N. Mashita, y T. Nishi. Three-dimensional structure of a nanocomposite material consisting of two kinds of nanofillers and rubbery matrix studied by transmission electron microtomography. *Macromolecules*, 40(18):6758–6764, 2007.

[Kas80] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 29(2):119–127, 1980.

[Kay84] B. H. Kaye. Fractal description of fineparticle systems. *Particle Characterization in Technology: Morphological analysis*, página 81, 1984.

[Ken83] J. T. Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[Kir71] R. A. Kirsch. Computer determination of the constituent structure of biological images. *Computers and biomedical research*, 4(3):315–328, 1971.

[Kit78] J. Kittler. Feature set search algorithms. *Pattern recognition and signal processing*, 41:60, 1978.

[KKI08] S. Kohjiya, A. Kato, y Y. Ikeda. Visualization of nanostructure of soft matter by 3D-TEM: Nanoparticles in a natural rubber matrix. *Progress in Polymer Science*, 33(10):979–997, 2008.

[KL04] R. B. Kale y C. D. Lokhande. Influence of air annealing on the structural, optical and electrical properties of chemically deposited CdSe nano-crystallites. *Applied surface science*, 223(4):343–351, 2004.

[KO11] R. Künzel y E. Okuno. X-ray spectroscopy applied to the study of the radiation transmission through nanomaterials. *Revista Brasileira de Física Médica*, 5(2):209–12, 2011.

[Kon94] I. Kononenko. Estimating attributes: analysis and extensions of relief. En *Machine Learning: ECML-94*, páginas 171–182. Springer, 1994.

[KP04] S. Kotsiantis y P. E. Pintelas. Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, 1(1):73–81, 2004.

[KPH07] A. Kalousis, J. Prados, y M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[KPK+12] K. K. Kandaswamy, G. Pugalenthi, K. U. Kalies, E. Hartmann, y T. Martinetz. EcmPred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. *Journal of theoretical biology*, 2012.

[KR92] K. Kira y L. A. Rendell. A practical approach to feature selection. En *Proceedings of the 9^{th} international workshop on Machine learning*, páginas 249–256. Morgan Kaufmann Publishers, 1992.

[KS96] D. Koller y M. Sahami. Toward Optimal Feature Selection. En *Proceedings of the International Conference on Machine Learning*, páginas 284–292. Citeseer, 1996.

[KSW85] J. N. Kapur, P. K. Sahoo, y A. K. C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer vision, graphics and image processing*, 29(3):273–285, 1985.

[KUH+07] O. P. Kocaoglu, S. R. Uhlhorn, E. Hernandez, R. A. Juarez, R. Will, J. M. Parel, y F. Manns. Simultaneous fundus imaging and optical coherence tomography of the mouse retina. *Investigative ophthalmology & visual science*, 48(3):1283–1289, 2007.

[LC87] W. E. Lorensen y H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. En *Proceedings of the 14^{th} annual conference on Computer graphics and interactive techniques*, página 169. ACM, 1987.

[LdGO+10] J. López-de-Uralde, T. Guraya, A. Okariz, E. Gómez, A. Zubillaga, y I. Santos. Automated Image Analysis and Artificial Intelligence applied to microscopic iamges: a new methodology for the morphological characterization of nano-particles. En *Proceedings of Modelling of Elastomeric Materials and Products*. 2010.

[LdRS+10] J. López-de-Uralde, I. Ruiz, I. Santos, A. Zubillaga, P. G. Bringas, A. Okariz, y T. Guraya. Automatic Morphological Categorisation of Carbon Black Nano-aggregates. *Lecture Notes in Computer Science*, 6262:185–193, 2010.

[Leb10] J. L. Leblanc. *Filled polymers: science and industrial applications*. CRC Press, 2010.

[Lew98] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *European Conference on Machine Learning (ECML'98)*, páginas 4–15, 1998.

[LG93] T. Lindeberg y J. Garding. Shape from texture from a multiscale perspective. En *Proceedings of the 4^{th} International Conference on Computer Vision*, páginas 683–691. IEEE, 1993.

[LGC97] L. Li, J. Gong, y W. Chen. Gray-level image thresholding based on Fisher linear projection of two-dimensional histogram. *Pattern Recognition*, 30(5):743–750, 1997.

[Lin93] T. Lindeberg. *Scale-space theory in computer vision*. Kluwer Academic Print on Demand, 1993.

[LL93] C. H. Li y C. K. Lee. Minimum cross entropy thresholding. *Pattern Recognition*, 26(4):617–625, 1993.

[LL98] C. K. Leung y F. K. Lam. Maximum segmented image information thresholding. *Graphical Models and Image Processing*, 60(1):57–76, 1998.

[Llo85] D. E. Lloyd. Automatic target classification using moment invariant of image shapes. *Report RAE IDN AW126, Farnborough, UK*, 1985.

[LLS92] L. Lam, S. W. Lee, y C. Y. Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 14(9):869–885, 1992.

[Low91] A. Low. *Introductory computer vision and image processing*. McGraw-hill, 1991.

[LP90] H. Lee y R. H. Park. Comments on an optimal threshold scheme for image segmentation. *IEEE Transactions on Systems, Man and Cybernetics*, 20:741–742, 1990.

[LT98] C. H. Li y P. K. S. Tam. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters*, 19(8):771–776, 1998.

[LY08] J. H. Lee y S. I. Yoo. An effective image segmentation technique for the SEM image. *IEEE International Conference on Industrial Technology (ICIT'08)*, 2008.

[Man77] B. B. Mandelbrot. Form, chance, and Dimension. *Chance and Dimension. Freeman, San Francisco*, páginas 1–234, 1977.

[Man82] B. B. Mandelbrot. *The fractal geometry of nature*. W.H. Freeman, San Francisco, 1982.

[Mar04] H. F. Mark. *Encyclopedia of Polymer Science and Technology*. John Wiley & Sons, 3ª edición, 2004.

[Mas97] D. N. Mastronarde. Dual-axis tomography: an approach with alignment methods that preserve resolution. *Journal of structural biology*, 120(3):343–352, 1997.

[MBM08] S. Maji, A. C. Berg, y J. Malik. Classification using intersection kernel support vector machines is efficient. En *IEEE Conference on Computer Vision and Pattern Recognition*, tomo 2, páginas 1–8. IEEE, 2008.

[McC07] J. McCarthy. What is artificial intelligence? Disponible en: `http://www-formal.stanford.edu/jmc/whatisai.pdf`, 2007.

[Mea83] P. Meakin. Formation of fractal clusters and networks by irreversible diffusion-limited aggregation. *Physical Review Letters*, 51(13):1119–1122, 1983.

[Med71] A. I. Medalia. Dynamic shape factors of particles. *Powder Technology*, 4(3):117–138, 1971.

[MEE05] J. E. Mark, B. Erman, y F. R. Eirich. *Science and technology of rubber*. Academic Press, 2005.

[Met78] C. E. Metz. Basic principles of ROC analysis. En *Seminars in nuclear medicine*, tomo 8, páginas 283–298. Elsevier, 1978.

[MG99] S. Maas y W. Gronski. Characterization of carbon blacks by transmission electron microscopy and advanced image analysis. *Kautschuk und Gummi, Kunststoffe*, 52(1):26–31, 1999.

[MGLLP+10] O. d. F. Martins-Gomes, P. R. Lopes-Lima, S. Paciornik, D. F. Brisola, y B. M. Cunha. Classification of fine particles from construction and demolition waste through image analysis. En *17th International Conference on Systems, Signals and Image Processing (IWSSIP'10)*, páginas 368–371. 2010.

[MGP05] O. F. Martins-Gomes y S. Paciornik. Automatic classification of graphite in cast iron. *Microscopy and Microanalysis*, 11(4):363, 2005.

[MH72] A. I. Medalia y G. J. Hornik. Pattern recognition problems in the study of carbon black. *Pattern Recognition*, 4(2):155–172, 1972.

[MLH12] I. A. Morozov, B. Lauke, y G. Heinrich. Quantitative Microstructural Investigation of Carbon-Black-Filled Rubbers by AFM. *Rubber Chemistry and Technology*, 85(2):244–263, 2012.

[MM73] J. N. Morgan y R. C. Messenger. *THAID: A sequential analysis program for the analysis of nominal scale dependent variables*. Survey Research Center, Institute for Social Research, University of Michigan, 1973.

[Mor87] M. Morton. *Rubber technology*. Chapman & Hall, 1987.

[MP90a] J. Malik y P. Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of Optical Society of America*, 7(5):923–932, 1990.

[MP90b] C. A. Murthy y S. K. Pal. Fuzzy thresholding: mathematical framework, bound functions and weighted moving average technique. *Pattern Recognition Letters*, 11(3):197–206, 1990.

[MR97] J. Malik y R. Rosenholtz. Computing local surface orientation and shape from texture for curved surfaces. *International Journal of Computer Vision*, 23(2):149–168, 1997.

[MU49] N. Metropolis y S. Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[MWW02] A. Maki, M. Watanabe, y C. Wiles. Geotensity: Combining Motion and Lighting for 3D Surface Reconstruction. *International Journal of Computer Vision*, 48(2):75–90, 2002.

[Nan12] Institute Of Nanotechnology. Nanotechnology Images. Disponible en: `http://www.nano.org.uk/nanotechnology-images`, 2012.

[NIK07] T. Nagai, M. Ikehara, y A. Kurematsu. Hmm-based surface reconstruction from single images. *Systems and Computers in Japan*, 38(11):80–89, 2007.

[NM10] T. L. Noguera Moreno. *Metodología ROC en la Evaluación de Medidas Antropométricas como Marcadores de la Hipertensión Arterial. Aplicación a Población Gallega Adulta*. Proyecto Fin de Carrera, Universidad de Santiago de Compostela, 2010.

[Nob79] O. Nobuyuki. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1), 1979.

[NSP+09] J. Nieves, I. Santos, Y. K. Penya, S. Rojas, M. Salazar, y P. G. Bringas. Mechanical Properties Prediction in High-Precision Foundry Production. En *Proceedings of the 7^{th} IEEE International Conference on Industrial Informatics (INDIN)*, páginas 31–36. 2009.

[OGL+10] A. Okariz, T. Guraya, I. Lecubarri, J. Eguzkitza, F. Garrido, y E. Gómez. 3D characterization techniques and Monte Carlo simulations to generate models of filled rubber. En *Proceedings of Modelling of Elastomeric Materials and Products*. 2010.

[Oli94] J. C. Olivo. Automatic threshold selection using the wavelet transform. *CVGIP: Graphical Models and Image Processing*, 56(3):205–218, 1994.

[OS97] C. Oh y C. M. Sorensen. The effect of overlap between monomers on the determination of fractal cluster morphology. *Journal of colloid and interface science*, 193(1):17–25, 1997.

[Ots75] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296, 1975.

[PA96] A. Pikaz y A. Averbuch. Digital image thresholding, based on topological stable-state. *Pattern Recognition*, 29(5):829–843, 1996.

[Pav80] T. Pavlidis. A thinning algorithm for discrete binary images. *Computer Graphics and Image Processing*, 13(2):142–157, 1980.

[PCRR+04] M. Poch, J. Comas, I. Rodriguez-Roda, M. Sanchez-Marre, y U. Cortés. Designing and building real environmental decision support systems. *Environmental Modelling & Software*, 19(9):857–873, 2004.

[PD00] F. Provost y P. Domingos. Well-trained PETs: Improving probability estimation trees, 2000.

[PdlC07] G. Pajares y J. M. de la Cruz. *Visión por Computador*. Ra-Ma Publishers, 2007.

[PI97] M. Peura y J. Iivarinen. Efficiency of simple shape descriptors. En *Proceedings of the 3rd international workshop on visual form*, páginas 443–451. 1997.

[PM90] P. Perona y J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[PP89] N. R. Pal y S. K. Pal. Entropic thresholding. *Signal processing*, 16(2):97–108, 1989.

[PR00] J. Pearl y S. Russell. Bayesian networks. *Handbook of brain theory and neural networks*, 2000.

[Pra07] W. K. Pratt. Digital Image Processing. *John Willey*, 2007.

[Pre70] J. M. S. Prewitt. *Object enhancement and extraction*, tomo 75. Academic Press, New York, 1970.

[PSS86] P. W. Palumbo, P. Swaminathan, y S. N. Srihari. Document image binarization: Evaluation of algorithms. En *Applications if Digital Image Processing IX*, páginas 278–286. International Society for Optics and Photonics, 1986.

[Qui79] J. R. Quinlan. Discovering rules by induction from large collections of examples. *Expert systems in the micro electronic age*, 174, 1979.

[Qui83] J. R. Quinlan. Learning efficient classification procedures and their application to chess end games. *Machine learning: An artificial intelligence approach*, 1:463–482, 1983.

[Qui86a] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[Qui86b] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[Qui93] J. R. Quinlan. *C4.5 programs for machine learning*. Morgan Kaufmann Publishers, 1993.

[Rad17] J. Radon. Berichte Sächsische Akademie der Wissenschaften. *Journal of Mathematical Physics*, página 262, 1917.

[RAS+00] K. Ramar, S. Arumugam, S. N. Sivanandam, L. Ganesan, y D. Manimegalai. Quantitative fuzzy measures for threshold selection. *Pattern Recognition Letters*, 21(1):1–7, 2000.

[RB06] M. Reisert y H. Burkhardt. Feature selection for retrieval purposes. *Image Analysis and Recognition*, páginas 661–672, 2006.

[RC78] T. W. Ridler y S. Calvard. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man and Cybernetics*, 8(8):630—-632, 1978.

[RdlT83] A. Rosenfeld y P. de la Torre. Histogram concavity analysis as an aid in threshold selection(in image processing). *IEEE Transactions on Systems, Man and Cybernetics*, 13(2):231–235, 1983.

[RN03] S. J. Russell y P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2ª edición, 2003.

[Rob65] L. G. Roberts. Machine Perception of Three-Dimensional Solids, in optical and Electro-Optical Information Processing. páginas 159–197, 1965.

[Rob77] G. S. Robinson. Edge detection by compass gradient masks. *Computer Graphics and Image Processing*, 6(5):492–501, 1977.

[RS06] E. Ribeiro y M. Shah. Computer vision for nanoscale imaging. *Machine Vision and Applications*, 17(3):147–162, 2006.

[RŠK03] M. Robnik-Šikonja y I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1):23–69, 2003.

[RT02] M. Rumpf y A. Telea. A continuous skeletonization method based on level sets. En *Proceedings of the symposium on Data Visualisation (VisSym'02)*, páginas 151–157. Eurographics Association, 2002.

[Rus87] J. C. Russ. Automatic discrimination of features in gray-scale images. *Journal of Microscopy*, 148(3):263—277, 1987.

[Rus07] J. C. Russ. *The image processing handbook*. CRC Press, 2007.

[RYS95] N. Ramesh, J. H. Yoo, y I. K. Sethi. Thresholding based on histogram approximation. *IEEE Proceedings - Vision, Image and Signal Processing*, 142(5):271–279, 1995.

[Sap01] G. Sapiro. *Geometric partial differential equations and image analysis*. Cambridge University Press, 2001.

[SAZL11] A. A. Sousa, A. A. Azari, G. Zhang, y R. D. Leapman. Dual-axis electron tomography of biological specimens: Extending the limits of specimen thickness with bright-field STEM imaging. *Journal of structural biology*, 174(1):107–114, 2011.

[SBH03] H. Scharr, M. J. Black, y H. W. Haussecker. Image statistics and anisotropic diffusion. En *Proceedings of the 9^{th} IEEE International Conference on Computer Vision (ICCV)*, tomo 2, páginas 840–847. IEEE Computer Society, 2003.

[SCN08] A. Saxena, S. H. Chung, y A. Y. Ng. 3-D Depth Reconstruction from a Single Still Image. *International Journal of Computer Vision*, 76(1):53–69, 2008.

[SDE+03] G. Schmid, M. Decker, H. Ernst, H. Fuchs, W. Grünwald, A. Grunwald, H. Hofmann, M. Mayor, W. Rathgeber, U. Simon, y D. Wyrwa. Small dimensions and material properties. *A Definition of Nanotechnology*, 11(3), 2003.

[Ser83] J. Serra. *Image analysis and mathematical morphology*. Academic Press, Inc. Orlando, FL, USA, 1983.

[Sez90] M. I. Sezan. A peak detection algorithm and its application to histogram-based image data reduction. *Computer vision, graphics and image processing*, 49(1):36–51, 1990.

[SF68] I. Sobel y G. Feldman. A 3x3 isotropic gradient operator for image processing. *Presentado en una charla en el Stanford Artificial Project*, páginas 271–272, 1968.

[SFF03] H. Scharr, M. Felsberg, y P. E. Forssén. Noise adaptive channel smoothing of low-dose images. En *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'03)*, tomo 2, páginas 18–18. IEEE Computer Society, 2003.

[SG92] S. C. Sahasrabudhe y K. S. D. Gupta. A valley-seeking threshold selection technique. *Computer vision and image processing*, 56:55–65, 1992.

[SG11] C. V. Sindelar y N. Grigorieff. An adaptation of the wiener filter suitable for analyzing images of isolated single particles. *Journal of structural biology*, 176(1):60–74, 2011.

[Sha94] A. G. Shanbhag. Utilization of information measure as a means of image thresholding. *CVGIP: Graphical Models and Image Processing*, 56(5):414–419, 1994.

[SHvT+01] T. Sorahan, L. Hamilton, M. van Tongeren, K. Gardiner, y J. M. Harrington. A cohort mortality study of UK carbon black workers, 1951–1996. *American journal of industrial medicine*, 39(2):158–170, 2001.

[SI97] D. Shen y H. H. S. Ip. A Hopfield neural network for adaptive image segmentation: an active surface paradigm. *Pattern Recognition Letters*, 18(1):37–48, 1997.

[SIL07] Y. Saeys, I. Inza, y P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[SKM09] Y. Singh, A. Kaur, y R. Malhotra. Comparative analysis of regression and machine learning methods for predicting fault proneness models. *International Journal of Computer Applications in Technology*, 35(2):183–193, 2009.

[SLTW04] V. Svetnik, A. Liaw, C. Tong, y T. Wang. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *Multiple Classifier Systems*, 5:334–343, 2004.

[SM09] N. F. Steinmetz y M. Manchester. PEGylated viral nanoparticles for biomedicine: the impact of PEG chain length on VNP cell interactions in vitro and ex vivo. *Biomacromolecules*, 10(4):784–792, 2009.

[SM11] M. Sankari y C. Meena. Object Matching using Skeletonization based on Hamming Distance. *International Journal of Computer Applications*, 28(7):46–50, 2011.

[SNPB09] I. Santos, J. Nieves, Y.K. Penya, y P.G. Bringas. Optimising Machine-learning-based Fault Prediction in Foundry Production. En *Proceedings of the 2nd International Symposium on Distributed Computing and Artificial Intelligence (DCAI)*. 2009.

[SS02] D. Scharstein y R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[SS04] M. Sezgin y B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168, 2004.

[SSA+91] M. Sumita, K. Sakata, S. Asai, K. Miyasaka, y H. Nakagawa. Dispersion of fillers and the electrical conductivity of polymer blends filled with carbon black. *Polymer bulletin*, 25(2):265–271, 1991.

[SWKN11] C. Som, P. Wick, H. Krug, y B. Nowack. Environmental and health effects of nanomaterials in nanotextiles and façade coatings. *Environment international*, 2011.

[SXM08] Z. Saghi, X. Xu, y G. Möbus. Three-dimensional metrology and fractal analysis of dendritic nanostructures. *Physical Review B*, 78(20):205428, 2008.

[TB10] D. Tschumperlé y L. Brun. Non-local image smoothing by applying anisotropic diffusion PDE's in the space of patches. En *16th IEEE International Conference on Image Processing (ICIP'10)*, páginas 2957–2960. IEEE, 2010.

[TC01] S. Tong y E. Chang. Support vector machine active learning for image retrieval. En *Proceedings of the 9th ACM international conference on Multimedia*, tomo 9, páginas 107–118. ACM, 2001.

[TDH11] R. Trappitsch, A. M. Davis, y P. R. Heck. Volume Measurement of Small Particles Using SEM Images. En *AGU Fall Meeting Abstracts*, tomo 1, página 1652. 2011.

[Tsa85] W. H. Tsai. Moment-preserving thresolding: A new approach. *Computer Vision, Graphics and Image Processing*, 29(3):377–393, 1985.

[TVW02] A. Telea y J. J. Van Wijk. An augmented fast marching method for computing skeletons and centerlines. En *Proceedings of the symposium on Data Visualisation, VisSym'02*, páginas 251–259. Eurographics Association, 2002.

[ÜMB07] B. Üstün, W. J. Melssen, y L. M. C. Buydens. Visualisation and interpretation of support vector regression models. *Analytica chimica acta*, 595(1-2):299–309, 2007.

[Vap99] V. N. Vapnik. The nature of statistical learning theory. Statistics for Engineering and Information Science, 1999.

[VC07] W. E. Vanderlinde y J. N. Caron. Blind Deconvolution of SEM Images. En *Proceedings of the 33rd International Symposium for Testing and Failure Analysis (ISTFA)*, tomo 33, páginas 97–102. ASM International, 2007.

[VCvSW86] O. Vohler, G. Collin, F. von Sturm, y E. Wege. Ullmann's encyclopedia of industrial chemistry. *VCH–New York*, 5:95–163, 1986.

[Vos86] R. Voss. Random fractals: Characterization and measurement. *Scaling Phenomena in Disordered Systems*, 133:1–11, 1986.

[VR95] S. Venkatesh y P. L. Rosin. Dynamic threshold determination by local and global edge evaluation. *Graphical Models and Image Processing*, 57(2):146–160, 1995.

[WCJ04] W. A. Wampler, T. F. Carlson, y W. J. Jones. Carbon Black. *Rubber Compounding*, páginas 239–284, 2004.

[Wei49] N. Weiner. *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*. MIT press, 1949.

[Wer12] Nano Werk. Introduction to Nanotechnology. Disponible en: `http://www.nanowerk.com/nanotechnology/introduction/introduction_to_nanotechnology_31.php`, 2012.

[WEST03] J. Weston, A. Elisseeff, B. Schölkopf, y M. Tipping. Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3(7-8):1439–1461, 2003.

[Wey02] M. Weyland. Electron tomography of catalysts. *Topics in catalysis*, 21(4):175–183, 2002.

[WFU06] A. Worster, J. Fan, y S. Upadhye. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine, CJEM*, 8(1):19–20, 2006.

[Wha91] R. J. Whatmough. Automatic threshold selection from a histogram using the exponential hull. *CVGIP: Graphical Models and Image Processing*, 53(6):592–600, 1991.

[Whi01] J. White. *Rubber technologist's handbook*, tomo 1. Rapra Technology, 2001.

[WN09] J. White y K. Naskar. *Rubber Technologist's Handbook*, tomo 2. Smithers Rapra Technology Limited, 2009.

[WR77a] J. S. Weszka y A. Rosenfeld. Histogram modification for threshold selection. *NASA STI/Recon Technical Report N*, 78, 1977.

[WR77b] J. S. Weszka y A. Rosenfeld. Threshold Evaluation Techniques. *Maryland University College Park Computer Science Center*, 1977.

[WR83] J. M. White y G. D. Rohrer. Image thresholding for optical character recognition and other applications requiring character image extraction. *IBM Journal of research and development*, 27(4):400–411, 1983.

[WWN+06] J. Wellmann, S. K. Weiland, G. Neiteler, G. Klein, y K. Straif. Cancer mortality in German carbon black workers 1976–98. *Occupational and environmental medicine*, 63(8):513–521, 2006.

[YB89] S. D. Yanowitz y A. M. Bruckstein. A new method for image segmentation. *Computer Vision, Graphics and Image Processing*, 46(1):82–95, 1989.

[YL04] L. Yu y H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[ZH08] J. Zhang y J. Hu. Image segmentation based on 2D Otsu method with histogram analysis. En *International Conference on Computer Science and Software Engineering*, tomo 6, páginas 105–108. IEEE, 2008.

[ZHH+09] S. Zhang, M. M. Hossain, M. R. Hassan, J. Bailey, y K. Ramamohanarao. Feature weighted SVMs using receiver operating

characteristics. En *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM09)*, páginas 497–508. 2009.

[ZTCS99] R. Zhang, P. S. Tsai, J. E. Cryer, y M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999.