Universidad de las Ciencias Informáticas

Facultad 2

Algoritmo de clasificación de nódulos pulmonares solitarios para alcanzar altos niveles de precisión

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autores: Jeffrey Artiles Lezcano

Luis Manuel Cruz Correa

Tutora: MSc. Arelys Rivero Castro

La Habana, 25 junio de 2015

"Año 57 de la Revolución"

"Una computadora merecería ser llamada inteligente si pudiera llevar a un ser humano a creer que no es humano."

Alan Turing

DECLARACIÓN DE AUTORÍA

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmamos la presente a los 25 días del mes de Junio del año 2015.

Autores:

_____ _____

Jeffrey Artiles Lezcano Luis Manuel Cruz Correa

Tutora:

_____

MSc. Arelys Rivero Castro

DATOS DE CONTACTO

Tutora:

MSc. Arelys Rivero Castro: graduada de ingeniera en Ciencias Informáticas, egresada de la UCI en el año 2009. En diciembre de 2014 defendió con resultados satisfactorios su tesis de maestría en Informática Aplicada. Ha impartido las asignaturas de Sistemas Operativos y Seguridad Informática. Se desempeñó como Asesora de Seguridad Informática en la Facultad 7. Actualmente se desempeña como Jefa de Proyecto en el Departamento de Desarrollo de Aplicaciones del Centro de Informática Médica de la Universidad de las Ciencias Informáticas.

Correo electrónico: arcastro@uci.cu.

DEDICATORIA

Dedicamos esta investigación a las personas que diariamente luchan contra el cáncer de pulmón. Que sepan que no están solos y que la comunidad científica está haciendo su mayor esfuerzo para mitigar los efectos de esta enfermedad.

AGRADECIMIENTOS

Al LuisMa por ser paciente, compañero, hermano y amigo.
Excepcionalmente a mi mamá, mi papá y a mis viejitos... sin ustedes no hubiera sido posible.
A Yoanys y Ariel por su preocupación permanente y apoyo en todo momento.
A mis hermanitas por tener que crecer conmigo lejos...
A mis Bisabuelos, abuelos, tíos y primos por apoyarme siempre.
A mi novia por ser como es.
En General a toda mi familia y amigos por estar pendiente a mis estudios.
Y especialmente a Arelys por confiar en nosotros y por tener la paciencia y la tranquilidad que siempre la caracteriza... una vez más GRACIAS!

Jeffrey

A Artiles por siempre confiar en mí y brindarme su amistad.

A mis padres por ser la guía de mis pasos y mi mayor aliento ante las adversidades, los quiero mucho.

A mi hermano, mi cuñada y mi sobrinita por apoyarme pese a la distancia.

A mi Crista por siempre estar ahí para mí y ser un ejemplo de constancia, el tiempo a tu lado ha sido maravilloso.

A mi abuelita quien a sus 90 años aún sigue siendo la primera persona que cada cumpleaños me felicita y me recuerda que la edad es solo una variable más, cuyo valor no afecta la ecuación del aprendizaje y el conocimiento.

A mis amigos, familiares y conocidos por darme el ánimo necesario en los momentos más complicados.

Y un especial agradecimiento al corazón de esta investigación, Arelys. No me alcanzan las palabras para agradecer la dedicación, la atención y la entrega para con nosotros. Ha sido un largo viaje, y este barco ha llegado a feliz puerto bajo la dirección de su capitana. Lo mucho que hemos aprendido durante la travesía se debe en gran parte a usted. Muchas Gracias!!!

Luisma

RESUMEN

El cáncer de pulmón se ha convertido en la primera causa de muerte oncológica en el mundo. Para minimizar el impacto de esta afección se han desarrollado sistemas de diagnóstico asistido por ordenador que emplean algoritmos para la clasificación de nódulos pulmonares solitarios en benignos o malignos. Dichos sistemas brindan a los radiólogos una segunda opinión en la interpretación de los resultados diagnósticos.

En la presente investigación se realiza un estudio de los algoritmos de clasificación que pudieran alcanzar mejores valores de precisión para problemas de este tipo. Se realiza un experimento comparando Red Neuronal, Máquina de Soporte Vectorial y k Vecinos más Cercanos; en el cual este último algoritmo arroja los mejores resultados de precisión. Se desarrolló un algoritmo de clasificación de nódulos pulmonares solitarios utilizando como Entorno de Desarrollo Integrado Visual Studio 2013 y como lenguaje de programación C# 4.0. Para realizar los experimentos y pruebas se empleó Matlab 2013 y Weka 3.7.10. Se utilizaron durante la fase de entrenamiento del clasificador, estructuras nodulares descritas en archivos XML asociados a estudios médicos, publicados en The Lung Image Database Consortium Image Collection. Se obtuvo un 81% de precisión en la clasificación de nódulos pulmonares solitarios, con una sensibilidad de 84% y una especificidad de 77%.

Palabras clave: algoritmo de clasificación, aprendizaje automatizado, nódulos pulmonares solitarios, precisión, vecinos más cercanos.

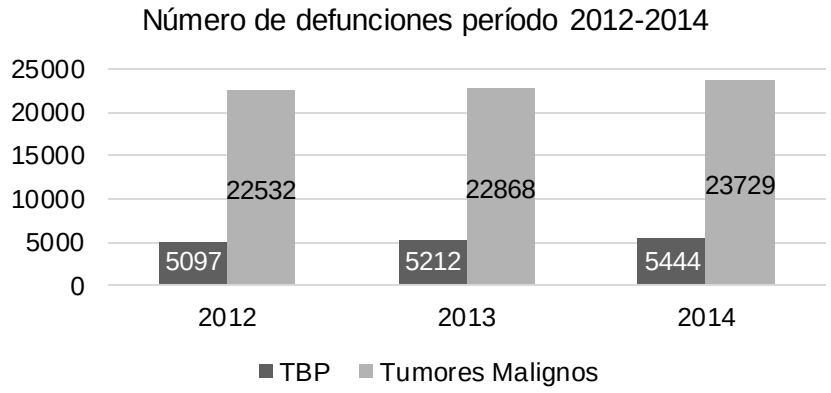
TABLA DE CONTENIDOS

INTRODUCCIÓN

El desarrollo de equipos médicos de alta tecnología ha permitido a los radiólogos realizar diagnósticos más rápidos y acertados. El trabajo con imágenes médicas ha posibilitado el surgimiento de una nueva rama de la medicina, la Imagenología. El diagnóstico por imágenes se utiliza con fines exploratorios, permite observar las estructuras internas del paciente. Las imágenes obtenidas permiten diagnosticar diferentes afecciones sin intervenir quirúrgicamente al paciente (1).

Existen diversas modalidades diagnósticas tales como la Resonancia Magnética (RM) (2), la Mamografía (MG) (3), la Radiografía Computarizada (RC) (4) y la Tomografía Computarizada (TC) (5). Esta última ha resultado especialmente útil para la identificación y clasificación de estructuras nodulares. Su principio de funcionamiento permite un alto nivel discriminatorio entre diferentes tipos de tejidos (6, 7). Debido a la resolución espacial y densidad o profundidad que presentan las imágenes de TC, se ha facilitado el estudio de diferentes afecciones y el análisis de enfermedades patológicas como es el caso del cáncer de pulmón (5, 8).

Entre los diferentes tipos de cáncer, el de pulmón constituye la primera causa de muerte en el varón y la tercera, después del de colón y mama, en la mujer (9). De acuerdo a datos de la Organización Mundial de la Salud (OMS), las muertes causadas por tumores malignos son las de mayor incidencia a nivel mundial (10). En Cuba se emite el Anuario Nacional Estadístico de Salud por el Ministerio de Salud Pública (MINSAP), en el cual se registran los datos relacionados con las más importantes causas de muerte ocurridas durante el año anterior. Los anuarios emitidos en los últimos tres años reflejan que los tumores malignos se han convertido en la primera causa de muerte en la isla. El mayor número de defunciones asociado a esta causa se localiza en la región tráquea, bronquios y pulmón (TBP). La Gráfica 1 muestra una distribución de estos valores durante este período.



***Gráfica 1**. Número de defunciones en el período 2012-2014 asociada a tumores malignos y específicamente al área TBP (11).*

El anuario emitido en el año 2015 (11) presenta una tasa de 48,8 defunciones por cada 100 000 habitantes; comportándose como el grupo de mayor prevalencia tanto en hombres como mujeres. La Tabla 1 recoge la distribución por género.

***Tabla 1.** Tasa de defunciones en Cuba debido a tumores malignos en tráquea bronquios y pulmón agrupada por género (11).*

Grupos	Defunciones por cáncer de TBP	Tasa por cada 100 000 habitantes
Hombres	3445	61.8
Mujeres	1999	35.7

Investigadores a nivel mundial se han dado a la tarea de estudiar la estructura y el funcionamiento de los sistemas de diagnóstico asistido por ordenador (CAD, por sus siglas en inglés), con el fin de ayudar a los especialistas en la toma de decisiones durante el proceso de diagnóstico del cáncer de pulmón. Los CAD tienen una estructura dividida en cuatro etapas: preprocesamiento, identificación de formas, reconocimiento de formas y clasificación (12). En la etapa de clasificación se realiza un proceso de discriminación a partir del cual, se logra disminuir considerablemente el número de falsos positivos. En esta fase es posible diferenciar las estructuras que son nódulos pulmonares solitarios de aquellas que no lo son. Los nódulos pulmonares solitarios son clasificados de acuerdo a su malignidad en benignos o malignos (13). El resultado del proceso proporciona una segunda opinión a los radiólogos, mejorando así el rendimiento y la eficacia del diagnóstico médico (12, 14).

En Cuba se han llevado a cabo investigaciones para el apoyo al diagnóstico médico. En la Universidad de las Ciencias Informáticas (UCI) se desarrolló un algoritmo de identificación de estructuras candidatas a ser nódulos pulmonares solitarios (15, 16). La incorporación de un clasificador a esta solución, posibilitaría mejorar la precisión en el resultado mostrado al especialista, al lograr una reducción importante del número de falsos positivos. En la fase de identificación de las estructuras nodulares a partir del análisis de las imágenes médicas de TC, existen un conjunto de limitaciones que inciden directamente en el correcto funcionamiento de la fase de clasificación, algunas de ellas son (17, 18):

- Los nódulos presentan un grupo de características (estructura interna, patrón de calcificación, esfericidad, bordes, lobulación, sutileza, textura y espiculación) que no siempre se manifiestan claramente en las imágenes y sus valores se encuentran solapadas entre la clasificación maligna y benigna.
- Elevado volumen de información clínica y radiográfica para ser analizada por el radiólogo.

Varias de las investigaciones llevadas a cabo por la comunidad científica internacional para el desarrollo de sistemas CAD (19, 20) emplean técnicas de inteligencia artificial (AI, por sus siglas en inglés) durante la etapa de clasificación. Entre las técnicas más utilizadas están las Redes Neuronales Artificiales (ANN, por sus siglas en inglés) (21) por su capacidad de hacer generalizaciones para resolver futuros problemas, Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) (22) por su facilidad de entrenamiento y k Vecinos Más Cercanos (kNN, por sus siglas en inglés) por sus fortalezas en la inducción ante datos ruidosos y funciones objetivo complejas (23, 24).

Las implementaciones de las técnicas de AI anteriormente mencionadas, presentan desventajas que impiden su generalización y aplicación en entornos reales. Los resultados obtenidos son aceptables pero es posible alcanzar mejores valores de precisión. Algunas de las principales desventajas que han sido identificadas por los autores de la investigación se encuentran relacionadas a continuación:

- Las implementaciones que utilizan kNN generalmente emplean el espectro completo de características del problema, sin realizar una distinción entre aquellas características que son redundantes, de las que son relevantes (25) y el rendimiento del clasificador se ve afectado.

- En algunas implementaciones que tienen como base una SVM no se identifican previamente los atributos relevantes para la construcción de la regla discriminante (26), por lo que los resultados obtenidos por el clasificador no son óptimos.
- Pocas implementaciones que hacen uso de la ANN realizan un preprocesamiento de los datos utilizados para el entrenamiento (19), por lo que la habilidad predictiva de la red se ve afectada por ruido en los elementos de entrada.

Por lo antes planteado se identifica como **problema de la investigación**: ¿Cómo alcanzar altos niveles de precisión en el proceso de clasificación de nódulos pulmonares solitarios?

El problema se enmarca en el **objeto de estudio**: proceso de clasificación de nódulos pulmonares solitarios empleando técnicas de inteligencia artificial, siendo el **campo de acción**: los algoritmos de clasificación supervisada para la clasificación de nódulos pulmonares solitarios.

Para dar solución al problema formulado se plantea como **objetivo general**: desarrollar un algoritmo de clasificación supervisada para alcanzar altos niveles de precisión en el proceso de clasificación de nódulos pulmonares solitarios.

Para dar solución al objetivo general planteado se proponen las siguientes **tareas de investigación**:

- Descripción de los algoritmos de clasificación de nódulos pulmonares solitarios desarrollados a nivel internacional y nacional.
- Determinación de los indicadores estadísticos que permiten medir el rendimiento del proceso de diagnóstico.
- Definición de los algoritmos de inteligencia artificial para la clasificación de estructuras nodulares.
- Desarrollo del algoritmo de clasificación de nódulos pulmonares solitarios.
- Validación de los resultados de precisión alcanzados por el algoritmo de clasificación de estructuras nodulares.

Los métodos científicos que se utilizan durante el desarrollo de la investigación son (27):

- Teóricos
    - **Analítico-Sintético**: al estudiar los diferentes elementos que conforman los algoritmos de clasificación de nódulos pulmonares solitarios; así como los requisitos

que deben cumplir una vez implementados. Determinar los elementos importantes a tener en cuenta durante la selección de los algoritmos de clasificación a utilizar.

- **Inductivo-Deductivo**: para analizar la problemática existente mediante la comprensión del funcionamiento del proceso de clasificación de estructuras nodulares en imágenes médicas y definir los elementos característicos, que permitan proponer un algoritmo que alcance altos valores de precisión en la clasificación.
- **Histórico-Lógico**: para investigar los antecedentes y evolución de los algoritmos de clasificación de estructuras nodulares. Permitió realizar el estudio del estado del arte de la problemática planteada mediante el análisis de diversas soluciones existentes e identificar las fuentes de información.
- **Modelación:** para representar gráficamente los diferentes elementos que componen el diseño del algoritmo y confeccionar los modelos y diagramas asociados al desarrollo del algoritmo.

- Empírico
  - **Experimento**: para determinar qué técnica de inteligencia artificial es la que mejores resultados arroja en la clasificación de nódulos pulmonares solitarios. Para validar los resultados obtenidos por el algoritmo de clasificación de nódulos pulmonares solitarios mediante la evaluación del indicador precisión.
- Estadísticos
  - **Estadística Descriptiva:** utilizada para recolectar las características de las estructuras nodulares, necesarias para realizar el proceso de clasificación.

Los resultados esperados con la puesta en práctica del algoritmo para la clasificación de nódulos pulmonares solitarios son:

- Elevar la precisión del proceso de clasificación de estructuras nodulares pudiendo contribuir a la reducción del tiempo de realización del diagnóstico médico.
- Contribuir en el proceso de formación de nuevos especialistas en el área de la Imagenología.

Estructura del contenido

Capítulo 1. Fundamentación teórica del algoritmo para la clasificación de nódulos pulmonares solitarios: reúne los principales conceptos vinculados a la investigación. Aborda un

estudio de las tendencias en el desarrollo de sistemas CAD enmarcándose en los algoritmos para la clasificación de nódulos pulmonares solitarios. Se describen los principales algoritmos de clasificación y se definen cuáles de ellos pueden ser relevantes para obtener resultados satisfactorios en la solución del problema planteado. Como parte de la revisión bibliográfica se detallan los pasos a seguir durante el proceso de clasificación de estructuras nodulares.

Capítulo 2. Características del algoritmo para la clasificación de nódulos pulmonares solitarios: describe cómo se realiza el flujo actual en las instituciones de salud para llevar a cabo el proceso de clasificación de estructuras nodulares y a partir de ello se conforma el Modelo de Dominio. Recoge los Requisitos Funcionales y no Funcionales a partir de los cuales se realiza una propuesta de solución para el problema planteado en la investigación. La propuesta de solución incluye la definición de los parámetros de configuración del algoritmo de clasificación. Se define la estrategia de trabajo con los datos que serán utilizados en el entrenamiento del algoritmo. Se modela el Caso de Uso del sistema relacionado con la clasificación de nódulos pulmonares solitarios.

Capítulo 3. Arquitectura, diseño, implementación y validación del algoritmo para la clasificación de nódulos pulmonares solitarios: aborda los elementos de diseño del algoritmo para la clasificación de nódulos pulmonares solitarios. Se conforma el Diagrama de Clases del Diseño y el Diagrama de Secuencia del Diseño. Se propone un modelo arquitectónico idóneo para el desarrollo del algoritmo referente a la presente investigación y se describen los patrones de diseño que se utilizan. Se describe la fase de implementación y se muestra el pseudocódigo del método más significativo del algoritmo propuesto para dar solución a la problemática planteada en la presente investigación. Se describen los indicadores a partir de los cuales se realizará la validación del algoritmo desarrollado para la clasificación de nódulos pulmonares solitarios. Queda evidenciado el cálculo de los indicadores y el análisis de los resultados.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA DEL ALGORITMO PARA LA CLASIFICACIÓN DE NÓDULOS PULMONARES SOLITARIOS

En este capítulo se describen los principales conceptos vinculados a la investigación así como las principales características morfológico-radiográficas de los nódulos pulmonares solitarios que son empleadas en la clasificación. Se definen los pasos que realiza un sistema de diagnóstico asistido por ordenador en su etapa de clasificación. Se incluyen los resultados del estudio de las tendencias del desarrollo de este tipo de sistemas, con el objetivo de identificar los pasos a seguir para la realización del algoritmo para la clasificación de nódulos pulmonares solitarios. Se analizan los diferentes algoritmos para la clasificación de estructuras nodulares para definir las características que debe tener el algoritmo que dará solución al problema planteado en la investigación.

1.1. Nódulo pulmonar solitario

Un nódulo pulmonar solitario (NPS) es aquella lesión única, redondeada, menor o igual de 30 mm de diámetro, que se encuentra rodeada completamente de parénquima[1] pulmonar normal, sin ninguna otra lesión acompañante (28). Los pacientes con NPS son por lo general asintomáticos, por lo que su detección ocurre en la mayoría de las ocasiones de forma casual (29).

La mayoría de los NPS son benignos, sin embargo pueden representar una etapa primaria del cáncer de pulmón. Los índices de supervivencia del cáncer de pulmón son muy bajos, apenas el 14% en un intervalo de 5 años. Si el tumor es detectado y clasificado cuando aún se encuentra en la fase de NPS, los índices de supervivencia se elevan hasta un 70-80%. Por lo que el diagnóstico temprano del NPS puede ser la única oportunidad de cura. (13)

1.2. Características morfológico-radiográficas de los nódulos pulmonares solitarios

Los NPS representados en las imágenes de Tomografía Computarizada (TC) presentan un grupo de características morfológico-radiográficas que son utilizadas para su clasificación. En la Tabla 2 se muestran estas características y los valores que pudieran tomar.

[1] Tejido conjuntivo esencial de ciertos órganos glandulares.

Tabla 2. *Valores que pueden tomar los nódulos pulmonares solitarios por cada una de sus características morfológico-radiográficas (30, 31).*

Características morfológico-radiográficas	Valores
sutileza	extremadamente sutil - obvio
estructura interna	tejido blando, fluido, grasa o aire
patrón de calcificación	palomita de maíz, laminada, sólida, no central, central o ausente
esfericidad	lineal, ovoide o redonda
bordes	pobremente definidos- bien definidos
lobulación	marcada o no lobulada
espiculación	marcada o no espiculado
textura	no sólida, vidrio deslustrado, parcialmente sólida o sólida

1.2.1. Criterios para la clasificación de nódulos pulmonares solitarios

Los criterios de clasificación permiten determinar la malignidad de los NPS con un alto grado de acierto. Estos criterios son utilizados por los especialistas para emitir un diagnóstico médico y son el resultado de análisis históricos realizados. Los criterios respecto a la clasificación de los NPS varían de un autor a otro, los más generalizados son (29):

Criterios que definen a un NPS como benigno

- Calcificación sólida difusa
- Forma redondeada, bordes lisos regulares, contenido graso, con/sin calcificación de palomita de maíz

Criterios que definen a un NPS como altamente sospechoso de malignidad (un solo criterio es suficiente)

- Textura de vidrio deslustrado mayor o igual a 10mm de diámetro
- Semisólidos
- Sólido mayor o igual a 20 mm de diámetro
- Sólido con contornos espiculados
- Sólido que contienen calcificaciones excéntricas o difusas

Criterios que definen un NPS como indeterminado

- Textura de vidrio deslustrado que mide menos de 10 mm de diámetro.
- Nódulo sólido menor de 20 mm de diámetro y contornos no espiculados

Estos criterios son aplicados por los radiólogos para emitir un diagnóstico. En muchas ocasiones la información que se extrae de las imágenes no es confiable, influyendo de manera negativa en el diagnóstico. En algunos casos es necesario recurrir a otros métodos más invasivos como biopsias o broncoscopias.

1.3. Diagnóstico asistido por computador

Como apoyo al diagnóstico médico, en la pasada década, ha sido extensivamente evaluado el potencial de los sistemas de diagnóstico asistido por ordenador (CAD, por sus siglas en inglés) para aumentar la habilidad de los radiólogos a la hora de detectar lesiones y específicamente clasificar NPS en TC de tórax (32, 33). Un sistema CAD realiza una caracterización de la información con el objetivo de detectar patrones en los datos y emitir un diagnóstico secundario al del experto radiólogo. Tiene como objetivo final ayudar a que el profesional mejore su rendimiento diagnóstico (12). En la Figura 1 se muestra de manera simplificada el flujo descrito anteriormente.

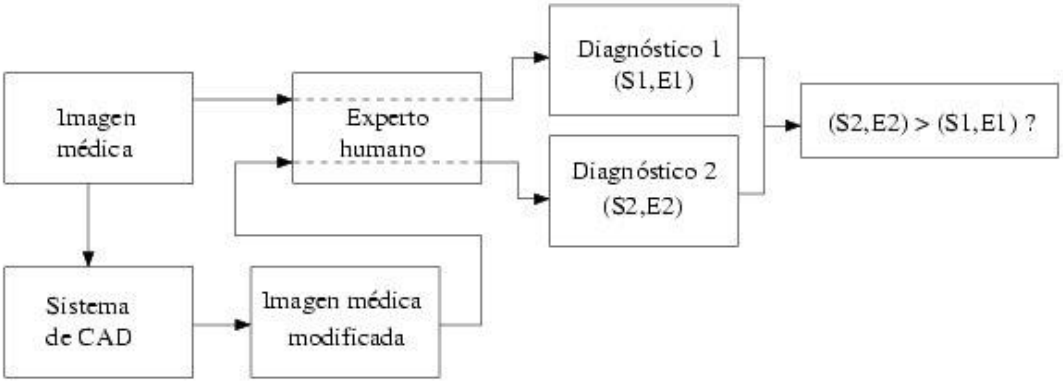


Figura 1. *Diagnóstico con y sin la ayuda de un sistema de CAD (4)*

Por las ventajas que presentan los sistemas CAD para incrementar el rendimiento diagnóstico de los radiólogos, su uso se ha incrementado en los últimos años como complemento de los sistemas de adquisición, archivado y comunicación de imágenes (PACS, por sus siglas en inglés) (12). La principal ventaja que conlleva esta integración es que además de las opciones de detección propias

de los PACS, los sistemas CAD permiten la integración de herramientas avanzadas de visualización y medición, que automáticamente permiten a los especialistas realizar una segmentación de un nódulo en 3D, un acercamiento volumétrico del tamaño y la densidad, o incluso la habilidad de identificar y clasificar nódulos pulmonares solitarios en series de imágenes de TC (34). La incorporación de estos dos sistemas a la misma estación de trabajo ha tomado vital importancia; un reporte de la *National Lung Screening Trial* (NLST) documenta una reducción del 20% en la demora diagnóstica al emplear la salida del CAD como una segunda opinión (35).

Los sistemas CAD cuentan con una estructura dividida en fases, cada fase tiene un objetivo en el proceso de transformación de la información inicial en conocimiento. La Figura 2 contiene una representación de cada fase.

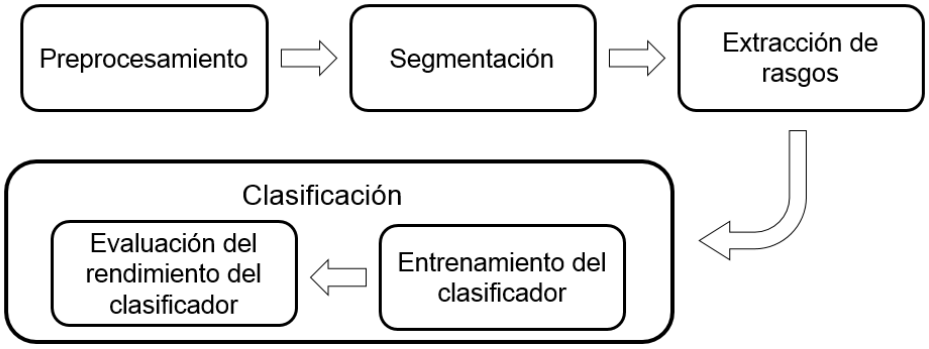


Figura 2. *Estructura interna de funcionamiento de un CAD (elaborada por los autores)*

La ***clasificación*** es la etapa en la que a partir de las características morfológicas o rasgos obtenidos se realiza una decisión sobre la clase a la que pertenecen los objetos. Los resultados obtenidos en la etapa de clasificación están influenciados por el entrenamiento que haya recibido el clasificador. El tipo de entrenamiento y la estructura del clasificador condicionan las características de la clasificación; por lo que su correcto diseño es relevante para el funcionamiento del CAD. La salida de esta etapa es evaluada en el análisis de rendimiento del clasificador, es en este momento donde se obtienen los indicadores que permiten evaluar el rendimiento del proceso diagnóstico (36). Esta etapa consta de dos fases implícitas: el entrenamiento del clasificador y la evaluación del rendimiento del clasificador (12).

El ***entrenamiento del clasificador*** evalúa cuáles son los valores de umbral que separan las clases, para esto se apoya en las reglas de decisión previamente definidas. Por lo general cuando se cuenta

con un conjunto de objetos previamente clasificados por algún método preciso, estos se utilizan como conjunto de entrenamiento del clasificador. (36)

La ***evaluación del rendimiento del clasificador*** da una medida de cuán bueno es el algoritmo clasificando nuevas instancias de objetos. La precisión del clasificador puede ser estimada tabulando su rendimiento sobre un conjunto de objetos de prueba. Para ello el conjunto debe ser representativo, y estar libre de errores. (36)

1.4. Tendencias en el desarrollo de la fase de clasificación de sistemas CAD enmarcados en el cáncer de pulmón

Relacionados con el desarrollo de sistemas CAD para la clasificación de nódulos pulmonares solitarios se han publicado varios artículos científicos que recogen las tendencias actuales en el diseño, desarrollo y empleo de algoritmos de clasificación como son:

En (37) seleccionan de la población de estudios publicados en *The Lung Image Database Consortium Image Collection* (LIDC), una muestra de 2000 imágenes de tomografía diagnosticadas. De estas imágenes se tomó una muestra aleatoria de 140 nódulos para formar conjuntos de entrenamiento para el modelo de red neuronal utilizado. Los restantes 60 nódulos, formando grupos de prueba, fueron presentados a la red ya entrenada y el rendimiento diagnóstico obtenido fue evaluado mediante un análisis de tipo *Receiver Operating Characteristic* (ROC) (38). El modelo de red neuronal artificial adoptado fue una red de retroalimentación con un algoritmo de propagación trasera, dividida en tres capas. La capa de entrada consistía en 12 neuronas y la capa oculta contaba con 7 neuronas decididas de forma experimental. La capa de salida tenía una neurona, cuyo valor de salida variaba de 0 a 1. La red fue entrenada usando la regla de Levenberg-Marquardt (39) y el entrenamiento fue terminado cuando la suma cuadrada del error fue menor que 0.001. Los resultados de la prueba demostraron que la red para los datos utilizados obtuvo una sensibilidad del 97.6%, especificidad del 84.2% y una precisión promedio del 93.3%.

En (40) se realizó un estudio con el fin de utilizar una red neuronal artificial para diferenciar entre nódulos pulmonares solitarios benignos y malignos, en imágenes de tomografía de alta resolución. Para el estudio se seleccionaron 155 nódulos menores de 3 cm (99 malignos y 56 benignos). Se utilizaron los parámetros clínicos (edad, sexo, historia como fumador, pérdida de peso y otros) y las características radiológicas (tamaño, forma, bordes, espiculación, calcificación y otros) extraídas por

radiólogos. Su rendimiento diagnóstico fue evaluado utilizando el análisis ROC y el índice Az que representa el área bajo la curva (AUC, por sus siglas en inglés). La red fue configurada utilizando un modelo de 3 capas, con retroalimentación y un algoritmo de propagación trasera. Se utilizaron 23 neuronas en la capa de entrada, cada una representando un parámetro clínico o una característica radiológica. Como resultado se obtuvo que la red neuronal mostró un alto rendimiento diferenciando los nódulos con un índice de Az=0.951.

En (41) se llevó a cabo un estudio con el objetivo de clasificar nódulos pulmonares solitarios en un conjunto desbalanceado de elementos candidatos usando *Support Vector Machine* (SVM). En el experimento se emplearon diferentes *kernels*[2] y parámetros de configuración para las SVM y diferentes conjuntos de entrenamiento balanceados obteniendo variados resultados. Los parámetros empleados para evaluar el rendimiento del clasificador fueron la sensibilidad y la especificidad. El propósito era alcanzar altos valores de sensibilidad para poder detectar todos los ejemplos positivos sin una pérdida considerable en especificidad. A los datos se le aplicaron SVM con *kernels* lineal, polinomial y gaussiano, variando el parámetro de regularización (C) entre 0.001 y 1000, los grados del *kernel* polinomial variaban de 2 a 6. Se calculó el error promedio y la desviación estándar en los conjuntos de entrenamiento y pruebas, así como la sensibilidad y la especificidad. Los valores obtenidos en cada uno de los modelos se evidencian en la Tabla 3:

Tabla 3. *Mejores modelos SVM con respecto a la sensibilidad (41).*

SVM *kernel*	Sensibilidad	Especificidad
Polinomial d=3,C=200	0.5333	0.9273
Polinomial d=2,C=1000	0.5317	0.9237
Polinomial d=4,C=1000	0.5317	0.8807
Polinomial d=6,C=200	0.53	0.89
Gauss. C=1000	0.53	0.9

[2] Función discriminante o de similaridad.

En (42) los autores llegaron a la conclusión de que los métodos tradicionales para la clasificación de nódulos pulmonares solitarios necesitaban extraer las características de las regiones de interés (ROI, por sus siglas en inglés). Generalmente este factor implica pérdida de información estructural implícita. Por tal motivo, con el objetivo de darle solución a la anterior problemática, aplicaron una novedosa técnica llamada *Multiple Kernel Learning method based on Matrix Least Square Support Vector Machine* (MKL-MatLSSVM). Para demostrar su efectividad fue aplicado en 20 estudios de diferentes pacientes, en los mismos se extrajeron regiones de interés que contenían 80 nódulos y 190 falsos positivos. Los resultados arrojaron que cuando se utilizan junto a esta técnica, *kernels* híbridos o *Radial Basis Function* (RBF), la sensibilidad, especificidad y precisión de la clasificación es balanceada. El área bajo la curva ROC puede alcanzar el 96%, valor notablemente superior que los obtenidos empleando otros métodos.

En (43) los autores implementan un CAD con el propósito de detectar y clasificar estructuras nodulares pulmonares, para el mismo se definen cada una de las fases del proceso. La etapa de clasificación emplea como algoritmo de clasificación *Clustering-K-Nearest-Neighbor*. Los datos utilizados pertenecen al conjunto estándar de la Sociedad de Tecnología Radiológica de Japón (JSRT, por sus siglas en inglés). El grupo de ejemplos tomados está compuesto por 154 nódulos de los cuales 100 son malignos y 54 benignos. La propuesta combina 2 algoritmos, el *k Means clustering* y el kNN con el objetivo de aumentar la precisión. Como resultado del estudio se obtuvo un valor de precisión del 98.7% lo que sustenta las capacidades de este algoritmo como clasificador.

En (17) se implementó un modelo del algoritmo kNN para efectuar la clasificación de una muestra compuesta por 400 nódulos. Fue utilizada la distancia euclidiana para realizar la asignación de clases y como medidor de desempate el más cercano. Se obtuvo como resultado una sensibilidad del 88% y una especificidad del 87%.

En (18) se emplea como algoritmo de clasificación *Fuzzy KNN*, una variante de kNN empleando lógica difusa, con el objetivo de clasificar nódulos pulmonares solitarios en imágenes 3D de tomografía computarizada. Para probar el método seleccionado se tomó una muestra de 63 estudios de la base de datos LIDC. El experimento demostró que el algoritmo alcanzó una sensibilidad del 88%, valor ligeramente superior al alcanzado por los radiólogos para el mismo conjunto de imágenes. Teniendo en cuenta que los nódulos eran de tamaño inferior a los 4mm y su forma era irregular, los autores llegaron a la conclusión de que los resultados obtenidos eran razonables.

1.5. Resultados de la investigación sobre las tendencias en el desarrollo de la fase de clasificación de sistemas CAD enmarcados en el cáncer de pulmón

Una vez realizado el análisis de las tendencias actuales en la investigación e implementación de algoritmos para la clasificación de nódulos pulmonares solitarios se han obtenido varias conclusiones:

1. Existen tendencias al empleo de los algoritmos Redes Neuronales Artificiales (ANN, por sus siglas en inglés), Máquina de Soporte Vectorial (SVM, por sus siglas en inglés) y k Vecinos Más Cercanos (kNN, por sus siglas en inglés) para la clasificación de nódulos pulmonares solitarios.
2. La elección del clasificador o regla de decisión depende de la naturaleza particular de los datos de entrada y de la salida esperada.
3. Algunos autores se enfocan en la obtención de altos valores de sensibilidad, sin prestar especial énfasis en la especificidad, cuando ambos indicadores inciden sobre los índices de precisión.

1.6. Descripción de los algoritmos para clasificar nódulos pulmonares solitarios identificados en el análisis de tendencias

A continuación se presenta una descripción simplificada del funcionamiento de los algoritmos k Vecinos Más Cercanos, Red Neuronal Artificial y Máquina de Soporte Vectorial. Se mencionan algunos de los parámetros que son necesarios tener en cuenta para su diseño e implementación. También se resaltan las principales ventajas y desventajas que condicionan el trabajo previo a realizar con los datos de entrenamiento.

1.6.1. k Vecinos Más Cercanos

Es un algoritmo de clasificación cuyo entrenamiento está basado en las instancias y características de los objetos que se analizan. En su funcionamiento, dicho algoritmo, ubica en un espacio n dimensional todas las instancias en forma de puntos, siendo el número de características la cantidad de dimensiones; de tal manera que los objetos que más cerca se encuentren van a ser más semejantes entre sí. La clasificación no comienza hasta que una nueva instancia u objeto arriba a ser clasificada. Entre las ventajas del algoritmo kNN resalta que es un método de inferencia inductiva

altamente efectivo cuando los datos presentan ruido o la función objetivo es compleja. La función objetivo para el espacio completo puede ser descrita como una combinación de aproximaciones locales menos complejas. Su entrenamiento es muy simple y cuenta con un indexado de memoria eficiente. Puede ser fácilmente integrado con sistemas de base de datos y aprovechar los métodos de acceso que estos proveen en forma de índices. (44)

Pese a sus ventajas, si este algoritmo se implementa de forma descuidada se hace necesario calcular todas las distancias entre los datos de entrenamiento y el dato de prueba; así como calcular los k vecinos más cercanos. Esto impacta en la escalabilidad del algoritmo de forma negativa. Aún con el empleo de métodos de acceso especializados, el costo de buscar los k vecinos más cercanos se incrementa significativamente a medida que se incrementa k (20). En la Figura 3 se aprecia la clasificación de un nuevo objeto para diferentes valores de k.

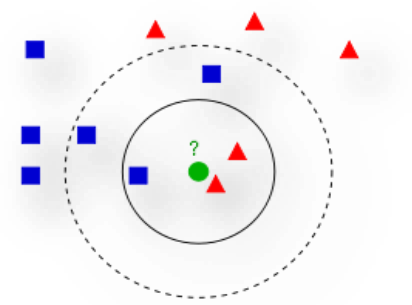


Figura 3. *Clasificación de una nueva instancia para k=3 da como resultado triángulo y k=5 da como resultado cuadrado (45).*

Dado un nuevo caso a clasificar, el algoritmo estima la salida teniendo en cuenta el tipo de problema que se está tratando de solucionar. Si el problema a resolver es de regresión (46), las predicciones se basan en el promedio de las salidas de los k vecinos escogidos. Si es un problema de clasificación, se emplea votación por mayoría y se obtiene como clase resultado la que cuente con más objetos semejantes (44).

La elección del parámetro k es uno de los factores más importantes y es esencial para la construcción del modelo. Su selección influye en la calidad de las predicciones. Pequeños valores de k pueden llevar a grandes variaciones en las predicciones; y por el contrario, grandes valores de k pueden llevar a la construcción de un modelo muy largo. Por tanto, k tiene que ser un valor lo suficientemente grande para minimizar la probabilidad de error en la clasificación y lo

suficientemente pequeño para que los k puntos más cercanos no estén muy alejados del punto de que se quiera evaluar.(45)

1.6.2. Redes Neuronales Artificiales

Las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano. Una red neuronal es "un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona".(47)

Las redes neuronales están compuestas de unidades de procesamiento que intercambian datos entre sí. Se utilizan para reconocer patrones en imágenes, manuscritos y secuencias de tiempo. A partir de un entrenamiento o aprendizaje inicial estas son capaces de inferir información y realizar tareas basadas en esos datos previamente almacenados. Por aprendizaje, se entiende que, la red puede "almacenar" la información adquirida por la modificación de los pesos de los enlaces de conexión entre las neuronas (19). Dada su capacidad de entrenamiento adaptativo, ante la aparición de nuevos casos pueden mejorar su funcionamiento y hacer generalizaciones para a partir de estos casos resolver futuros problemas.

Como modelo es especialmente utilizado para resolver casos en los cuales sea necesario realizar aproximaciones no lineales (48). Tiene como desventajas la naturaleza empírica del diseño de la red y la gran carga computacional que suponen para su ejecución (49). Su principal desventaja es, que en el entrenamiento pueden ser afectadas por ruido en los datos, lo que da como resultado que disminuya la habilidad predictiva de la red (50). Las condiciones que influencian este problema son: el tamaño de la red y el tiempo de entrenamiento. Si el tiempo de entrenamiento es excesivo ocurre un problema de sobreentrenamiento provocando que los resultados predictivos sean malos (51).

Con el objetivo de evitar el sobreentrenamiento en las redes neuronales se han aplicado con éxito dos técnicas (52):

- Detener el entrenamiento a tiempo: consiste en determinar el momento adecuado para detener el entrenamiento de la red, evitando la pérdida de generalización mediante el establecimiento del criterio de parada correcto.
- Incrementar el intervalo de entrenamiento: consiste en incrementar la cantidad de información con la cual se está entrenando la red

1.6.3. Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial son una especialización de las redes neuronales artificiales unidireccionales (48). Su funcionamiento se basa en el uso de un *kernel* o función discriminante encargado de inducir un "espacio de características", en el cual se realiza la separación y el traslado de las muestras. El algoritmo da la posibilidad de clasificar en múltiples clases, por lo que es fácilmente adaptable a muchos de los problemas de este tipo que se presentan en la actualidad. En su proceso se utilizan múltiples clasificadores binarios, donde cada uno aporta un voto a la clasificación, al final, se clasifica en el grupo que más votos haya tenido. El entrenamiento es relativamente fácil; pero una debilidad es que se necesita seleccionar una "buena" función *kernel* que permita sintonizar los parámetros de inicialización de la SVM de manera eficiente (22). Por defecto SVM es un clasificador binario pero si es necesario resolver un problema que tiene más de dos clases, es necesario construir tantos clasificadores como clases empleando una estrategia de **uno contra todos** (53).

1.7. Comparación de los algoritmos kNN, ANN y SVM

Una vez identificado que existe una tendencia en el uso de técnicas como kNN, ANN y SVM en el desarrollo de clasificadores para nódulos pulmonares solitarios, los autores de la presente investigación, analizaron los resultados obtenidos por otros investigadores, en estudios comparativos realizados entre estos tres algoritmos.

1.7.1. Comparación entre kNN y ANN

En (21) los autores realizaron un experimento para poder elaborar una comparación entre los algoritmos kNN y ANN. De dichos algoritmos, ANN es más complejo, porque dispone de varios parámetros que deben ser establecidos antes de diseñar el modelo de la red neuronal. El modelo, tamaño, función de activación, parámetros de aprendizaje y el número de muestras de entrenamientos se encuentran entre estos parámetros. Luego de realizada las pruebas de un diagnóstico de imágenes de espectrograma, se concluyó que kNN arrojaba mejores resultados en términos de especificidad.

En (23) realizaron un estudio cuyo objetivo es analizar las diferencias y similitudes en las dos reglas de aprendizaje supervisado y determinar si un clasificador es más adecuado para ciertos problemas de clasificación. Las pruebas realizadas demostraron que el conjunto de entrenamiento tiene una

gran influencia en las capacidades de entrenamiento del clasificador. El parámetro comparado en ambos algoritmos es la optimización y el experimento demostró que kNN presenta mejores resultados en sensibilidad y especificidad que ANN, para el mismo tamaño de muestra.

1.7.2. Comparación entre kNN y SVM

En (53) decidieron comparar versiones optimizadas de los algoritmos kNN y SVM. Los resultados mostraron que SVM a pesar de tener un buen rendimiento en general, no sobrepasaba a kNN en especificidad y precisión. Si un procesamiento previo de la información es empleado con kNN, el algoritmo mantiene altos valores de precisión incluso al ser aumentado el tamaño del conjunto de datos. No ocurre de la misma forma para la SVM que al aumentar la cantidad de datos, su tiempo de respuesta aumenta de forma cuadrática.

De igual forma en (24) se realiza un experimento basándose en el reconocimiento de patrones en un grupo de imágenes dividido en tres categorías: *modas, flores y personas africanas*. Este estudio compara el rendimiento de los algoritmos basado en el graficado de la curva de aprendizaje. Como resultado se obtiene que kNN presenta valores de precisión por encima del 90%, mientas que SVM presenta un comportamiento inestable, quedando por debajo de kNN en dos de las tres categorías de imágenes probadas.

Resultados del estudio comparativo entre kNN, ANN y SVM

Después de realizar un estudio de artículos que compararan los distintos algoritmos, los autores evidencian que existe diversidad de criterios y el origen de los datos es diferente. No es posible determinar qué algoritmo ofrece mejores resultados de precisión realizando una comparación directa entre los estudios. Para determinar cuál de las técnicas estudiadas pudiera arrojar mejores resultados de rendimiento al ser utilizada para clasificar nódulos pulmonares solitarios, los autores de la presente investigación deciden realizar un experimento, que emplee los mismos datos para entrenar y probar los algoritmos: y así igualar las condiciones. Los datos a utilizar serán obtenidos de una base de datos internacional de estudios médicos.

1.8. Elección del modelo de algoritmo de clasificación a utilizar

Para realizar el experimento es necesario definir el origen de los datos que serán utilizados en el entrenamiento y prueba de los algoritmos, así como sus características. A nivel internacional han

sido desarrolladas bases de datos (BD) que recopilan información asociadas a procesos médicos con el objetivo de potenciar el desarrollo de sistemas CAD. Estas BD reúnen los resultados de análisis llevados a cabo por especialistas y sirven de guía para el desarrollo de herramientas computacionales que utilicen estos datos para inferir nuevo conocimiento (54, 55). Para el desarrollo de CAD orientados a la clasificación de nódulos pulmonares solitarios se han destacado las siguientes BD:

- *The Lung Image Database Consortium Image Collection - Image Database Resource Initiative (LIDC/IDRI)* (56, 57).
- *Japanese Society of Radiological Technology (JSRT)* (43, 58).
- *Early Lung Cancer Action Program Public Lung Image Database (ELCAP)* (57).
- *CT Image Library (CTIL)* (59).

Existe un número significativo de investigaciones que hacen uso de la LIDC/IDRI por sus ventajas respecto a las otras BD mencionadas. Entre las ventajas se encuentran el mayor tamaño de la BD y que la misma contiene anexo a las series de imágenes de cada estudio, un fichero XML que describe según la norma del *National Institute of Health* (NIH) y el *The Cancer Imaging Archive* (TCIA) aprobada en el año 2006 y actualizada en el año 2010, las características más relevantes de las estructuras nodulares presentes en dichas imágenes (57, 60). En el XML se especifican los datos asociados al estudio en general y las descripciones de las estructuras nodulares encontradas por los radiólogos. Entre el conjunto de parámetros, el XML almacena la probabilidad de malignidad asociada a cada estructura nodular variando desde 1 hasta 5.

Empleando los datos contenidos en 30 series de imágenes, los autores de la presente investigación desarrollaron un experimento, para determinar cuál de los algoritmos SVM, ANN y kNN sería utilizado como modelo para la implementación del algoritmo para la clasificación de nódulos pulmonares solitarios. Para lograrlo, modelaron el funcionamiento de los algoritmos encontrados en ambiente Matlab. La implementación de los algoritmos en este asistente matemático permite modificar sus parámetros para simular la igualdad de condiciones. El conjunto de datos empleado en cada iteración de prueba fue el mismo para cada uno de los algoritmos. La prueba de rendimiento fue realizada utilizando una matriz de aprendizaje que contenía 68 nódulos, representando el 70% de los datos. La matriz de pruebas contenía 14 nódulos al igual que la matriz de validación, siendo el 15% de los datos en ambos casos.

De los algoritmos, la SVM fue empleada con el *kernel Radial Basis Function*. La ANN era un perceptrón multicapa entrenada siguiendo el algoritmo de Levenberg-Marquardt. El algoritmo kNN fue probado para distintos valores de k. Los resultados del experimento se evidencian en la Tabla 4. La prueba realizada arrojó como resultado que kNN con k=3 obtuvo mejores resultados en cuanto a precisión que los otros algoritmos, para el juego de datos utilizados en el experimento.

Tabla 4. *Resultados del experimento para determinar el modelo a seguir para la implementación del algoritmo de clasificación (elaborada por los autores).*

Algoritmo	Precisión promedio
kNN k=3	0.87856
kNN k=5	0.86427
RNA	0.81428
SVM	0.83571

1.9. Metodologías, tecnologías, herramientas y lenguajes a utilizar en el desarrollo del algoritmo de clasificación de nódulos pulmonares solitarios

Proceso Racional Unificado (RUP, por sus siglas en inglés): es una metodología robusta que provee un acercamiento disciplinado a la asignación de tareas y responsabilidades en una organización de desarrollo. Es utilizada para el análisis, implementación y documentación de sistemas. Los proyectos RUP son iterativos e incrementales y su progreso es medido por hitos. Su objetivo es la obtención de un software eficiente, mediante la planificación total del trabajo a realizar antes del comienzo del ciclo de desarrollo. RUP es una guía de cómo usar efectivamente el Lenguaje Unificado de Modelado (UML, por sus siglas en inglés) en la creación de los artefactos de Requerimiento, Arquitectura y Diseño. (61, 62)

Lenguaje Unificado de Modelado (UML, por sus siglas en inglés): se utiliza para la creación, documentación, especificación y visualización de los distintos artefactos que se generan en el desarrollo de un software. Emplea un esquema orientado a los objetos que incluye diagramas que proveen de múltiples perspectivas del sistema en análisis. Esta organizado en paquetes y

estructurado por capas, que permiten a los desarrolladores, crear diseños de forma fácil para ser compartidos con otros desarrolladores o clientes. (63, 64)

Visual Studio 2013: es un Entorno de Desarrollo Integrado (IDE, por sus siglas en inglés), que incluye un variado grupo de lenguajes de programación entre los que se encuentra Visual C#. El IDE cuenta con un editor de código que proporciona un completo *intellisense* predictivo y múltiples herramientas de refactorización y aceleración de la codificación. Además permite la prueba de las aplicaciones en cuanto a carga y rendimiento, brindando consejos para mejorar el código. Al ser un IDE extensible se le pueden incorporar mediante *plugins* nuevas funcionalidades que aumentan su potencial. (65)

C# 4.0: es un lenguaje de programación orientado a objetos que permite a los desarrolladores crear un código fácil de mantener. Esta versión incorpora tipos genéricos, clases parciales y otras construcciones de lenguaje que son útiles para generar código en menos tiempo y de fácil entendimiento. Su sintaxis permite el empleo de encapsulación, herencia y polimorfismo en la creación de aplicaciones. (66)

Enterprise Architect 7.5: es una herramienta para el desarrollo avanzado de modelado que cuenta con una interfaz intuitiva y un alto rendimiento. Esta herramienta cubre el desarrollo de software desde el paso de los Requerimientos a través de las etapas de Análisis, Modelos de diseño, Pruebas y Mantenimiento. Ofrece salida de documentación flexible y de alta calidad. Facilita realizar ingeniería inversa de código fuente para el lenguaje de programación C#. (67)

Tortoise SVN 1.7.6: es una herramienta de código abierto para el control de versiones tanto en documentos como en el código. Es intuitivo y fácil de usar, su uso es libre de costo. Puede ser fácilmente integrado a Visual Studio mediante plugins o mediante comandos. Emplea un repositorio centralizado que es capaz de gestionar los cambios que se hacen es sus directorios o en sus ficheros. (68)

Matlab 2013: es un poderoso asistente matemático interactivo utilizado para la computación numérica, la visualización y la programación. Cuenta con un lenguaje de programación de alto nivel que permite analizar datos, desarrollar algoritmos y crear modelos de aplicaciones. Las herramientas y las funciones matemáticas que incorpora permiten explorar múltiples enfoques en la resolución de problemas y alcanzar soluciones con mayor rapidez que otros lenguajes. Puede ser utilizado para

llevar a cabo experimentos en un ambiente controlado, permitiendo una total manipulación de las variables implicadas. (69)

Weka 3.7.10: es un software de código abierto desarrollado en JAVA y licenciado bajo Licencia Pública General (GPL, por sus siglas en inglés), este contiene una colección de algoritmos de aprendizaje automatizado para realizar tareas de minería de datos. Está dotado de un conjunto de herramientas para el preprocesamiento de los datos, su clasificación, análisis de regresión, clustering, establecer reglas de asociación o visualizar la información en forma de gráficos. Puede ser empleado para desarrollar nuevos esquemas de aprendizaje automatizado. Cuenta con módulos específicos para la realización de experimentos y el graficado del flujo de conocimiento en los algoritmos. (70)

Una vez analizadas las herramientas se emplearán el IDE Visual Studio 2013 y el lenguaje de programación C# 4.0, debido a sus funcionalidades para el desarrollo sencillo de algoritmos. Para el modelado del algoritmo de clasificación de nódulos pulmonares solitarios será utilizado el Enterprise Architect 7.5 y para el control de versiones el Tortoise SVN 1.7.6 debido a que ambos se integran fácilmente con Visual Studio. El empleo de las herramientas anteriormente mencionadas, está definido por el Centro de Informática Médica (CESIM) para el desarrollo de software en el proyecto PACS-RIS.

1.10. Modelo de Madurez de la Capacidad de Integración

Modelo de Madurez de la Capacidad de Integración (CMMI, por sus siglas en inglés) es un conjunto de modelos elaborados por el Instituto de Ingeniería del Software de Estados Unidos (SEI, por sus siglas en inglés) con el objetivo de obtener un diagnóstico preciso de la madurez de los procesos relacionados con las tecnologías de la información de una organización, y describen las tareas que se tienen que llevar a cabo para mejorar esos procesos (71). CMMI mide los niveles de madurez a través de calificaciones que reciben las organizaciones cuando son evaluadas (72), siendo estas:

1. Inicial
2. Administrado
3. Definido
4. Cuantitativamente administrado
5. Optimizado

El CESIM perteneciente a la UCI certificó en el año 2011 el nivel dos de CMMI. Actualmente el centro se encuentra en proceso de ratificación de dicho nivel, línea que seguirán los proyectos que en él se desarrollan.

1.11. Conclusiones del capítulo

Una vez realizado el análisis de los diferentes algoritmos empleados en la clasificación de estructuras nodulares, se arribaron a las siguientes conclusiones:

- Se identificó que los pasos básicos seguidos por los desarrolladores de algoritmos para la clasificación de nódulos pulmonares solitarios son: entrenamiento del clasificador, clasificación de estructuras nodulares y evaluación del rendimiento del clasificador.
- Existe una tendencia al empleo de base de datos internacionales para entrenar y validar los algoritmos de clasificación de nódulos pulmonares solitarios, la más relevante de ellas es la LIDC.
- Se determinó que en el marco de la investigación el algoritmo kNN arroja mejores resultados de precisión que otras técnicas de clasificación como SVM y ANN.

CAPÍTULO 2. CARACTERÍSTICAS DEL ALGORITMO PARA LA CLASIFICACIÓN DE NÓDULOS PULMONARES SOLITARIOS

En el presente capítulo se tiene como objetivo, explicar cómo lleva a cabo el especialista el proceso de clasificación de nódulos pulmonares solitarios. A partir del flujo descrito se realiza el Modelo de Dominio correspondiente, para luego extraer las características con las que contará el algoritmo a desarrollar. Se propone una solución al problema planteado para la investigación teniendo en cuenta el entrenamiento del clasificador. Se muestra el Diagrama de Casos de Uso del Sistema asociado al algoritmo.

2.1. Proceso de clasificación de nódulos pulmonares solitarios

Para realizar el proceso de clasificación de nódulos pulmonares solitarios el especialista previamente debe identificar dichas estructuras y determinar los valores de las características que los identifican. Las estructuras pulmonares pueden ser descritas teniendo en cuenta sutileza, esfericidad, patrón de calcificación, lobulación, estructura interna, bordes, espiculación y textura. Para realizar el análisis e interpretación de las características, el especialista hace uso de una estación de visualización de imágenes médicas y de su conocimiento tácito. El conjunto de descriptores resultante de la extracción de rasgos es la entrada del proceso de clasificación y posibilita que el especialista sea capaz de determinar cuáles de las estructuras son benignas o malignas.

2.2. Modelo de Dominio

El Proceso Racional Unificado (RUP, por sus siglas en inglés) con el fin de esclarecer el entorno del problema define en su primera fase de desarrollo la realización de un Modelo de Casos de Uso del Negocio. Cuando los procesos del negocio no se encuentran bien identificados, RUP propone realizar un Modelo de Dominio. El Modelo de Dominio tiene como objetivo comprender y describir solamente las clases más importantes dentro del contexto en el cual se desempeña el software, con el propósito de sentar las bases del entendimiento del desarrollo y no para definirlo completamente. (73)

Para realizar el proceso de clasificación de estructuras nodulares, el especialista emplea una estación de visualización que cuenta con un grupo de herramientas y funcionalidades. Estas le permiten extraer las características de las estructuras nodulares mediante su manipulación. Una vez

que se obtienen los descriptores de los posibles nódulos pulmonares solitarios, este procede a determinar si los mismos son efectivamente estructuras nodulares o no. Una vez aplicada la primera discriminación realiza el análisis para determinar la malignidad teniendo en cuenta los factores que indican malignidad o benignidad. El Modelo de Dominio aparece representado en la Figura 4.

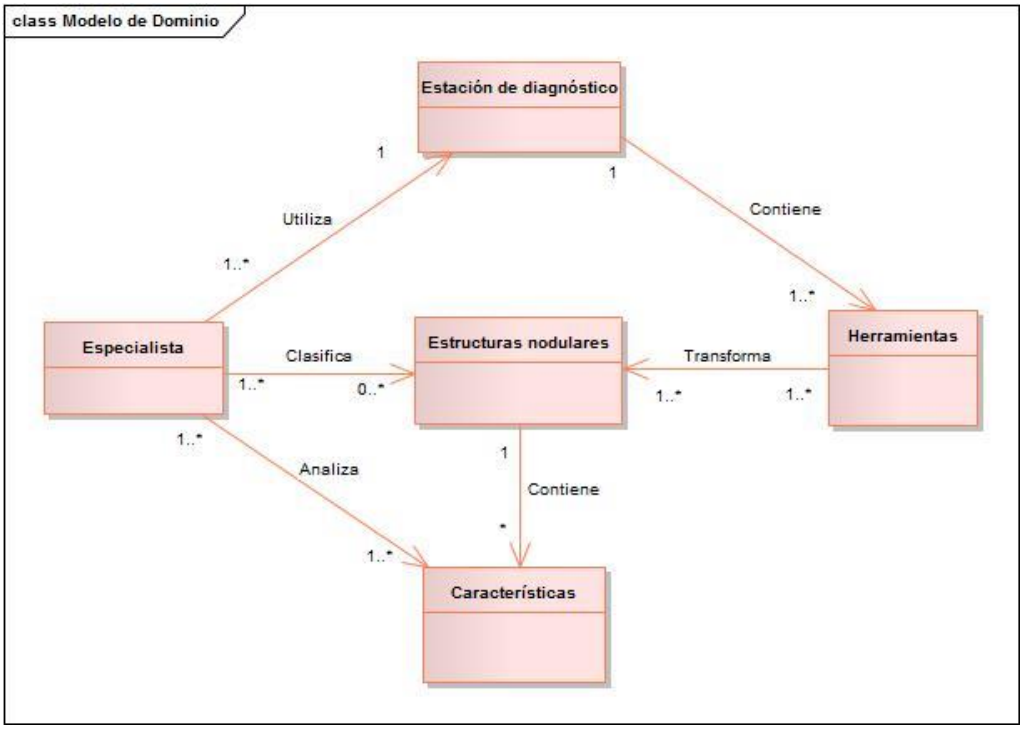


Figura 4. *Modelo de Dominio del algoritmo de clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

Para tener conocimiento acerca de la descripción de las entidades y conceptos que conforman el Modelo de Dominio ver la Tabla 5.

Tabla 5. *Entidades y conceptos fundamentales (elaborada por los autores)*

Entidades y conceptos	Descripción
Especialista	Radiólogo que mediante herramientas extrae las características de las estructuras nodulares y clasifica las estructuras.

Entidades y conceptos	Descripción
Estación de diagnóstico	Contiene las herramientas que permiten al radiólogo determinar el valor de las características de las estructuras nodulares.
Estructuras nodulares	Objetos que serán clasificados y a los cuales se les extraen sus características para realizar este proceso.
Características	Son los descriptores que se extraen de los nódulos pulmonares solitarios detectados, como pueden ser: sutileza, estructura interna, calcificación, esfericidad, bordes, espiculación, lobulación y textura.
Herramientas	Conjunto de funcionalidades y herramientas que le permiten al especialista manipular los parámetros de visualización de las estructuras nodulares. Contribuyen al proceso de extracción de características morfológico-radiográficas de las estructuras nodulares mediante su manipulación.

2.3. Requisitos Funcionales del algoritmo de clasificación de nódulos pulmonares solitarios

Los requisitos funcionales describen qué debería hacer el sistema y pueden ser expresados en términos de requisitos funcionales de usuario y de sistemas. Los primeros describen el sistema de forma abstracta, mientras que los segundos describen el funcionamiento del sistema en detalle, sus entradas y sus salidas (74). Los requisitos funcionales se muestran en la Tabla 6:

Tabla 6. *Requisitos Funcionales del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

Requisitos	Descripción
RF 1 Clasificar estructuras nodulares	Recibe las características morfológico-radiográficas de las estructuras nodulares en forma de vector

Requisitos	Descripción
	característico. Realiza el procesamiento y concluye arrojando el nivel de malignidad.
RF 1.1 Calcular distancia euclidiana.	Cálculo de la distancia euclidiana de la nueva instancia a clasificar respecto a las demás instancias previamente almacenadas.
RF 1.2 Seleccionar las k instancias más cercanas	Selecciona las k instancias más cercanas a la instancia a clasificar teniendo en cuenta el cálculo de las distancias.
RF 1.3 Calcular el valor de la función objetivo	Se calcula la función objetivo que devuelve la clasificación de la nueva instancia, teniendo en cuenta la estrategia de desempate en caso de existir alguno.

En la Figura 5 se muestra el Diagrama de Requerimientos Funcionales agrupados en forma de paquetes lógicos.

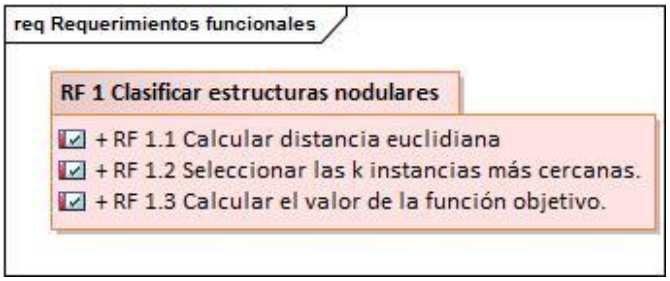


Figura 5. *Requisitos Funcionales para el algoritmo de clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

2.4. Requisitos no Funcionales

Los Requisitos no Funcionales no se refieren directamente a las funcionalidades desarrolladas por el sistema. Se relacionan con las características del sistema tales como confiabilidad, tiempo de respuesta y requerimientos técnicos. Definen restricciones sobre las propiedades que especifican el rendimiento del sistema, su disponibilidad y su organización. (74)

Los Requerimientos no Funcionales del algoritmo de clasificación de nódulos pulmonares solitarios aparecen a continuación:

Funcionamiento:

RNFO 1. Utilizar el sistema operativo Windows 7 o superior

RNFO 2. Utilizar 1 GB de RAM o superior para funcionar de forma óptima

RNFO 3. CPU Pentium IV 3.0GHz o superior.

Diseño e Implementación:

RNDI 1. Visual Studio 2013 como Entorno de Desarrollo Integrado

RNDI 2. C# como lenguaje de programación.

RNDI 3. Utilizar Enterprise Architect 7.5 como herramienta CASE.

RNDI 4. Uso de Framework.Net 4.5.

RNDI 5. Uso de UML como lenguaje de modelado.

RNDI 6. Estándar de codificación

RNDI 7. TortoiseSVN como herramienta para el control de versiones.

Legal:

RNL 1. El algoritmo para la clasificación de nódulos pulmonares solitarios es propiedad del Departamento de Desarrollo de Aplicaciones.

Usabilidad:

RNU 1. Se debe mostrar al usuario un mensaje con el valor de malignidad del nódulo que se está evaluando.

En la Figura 6 se muestra el diagrama de Requisitos no Funcionales agrupados por paquetes lógicos.

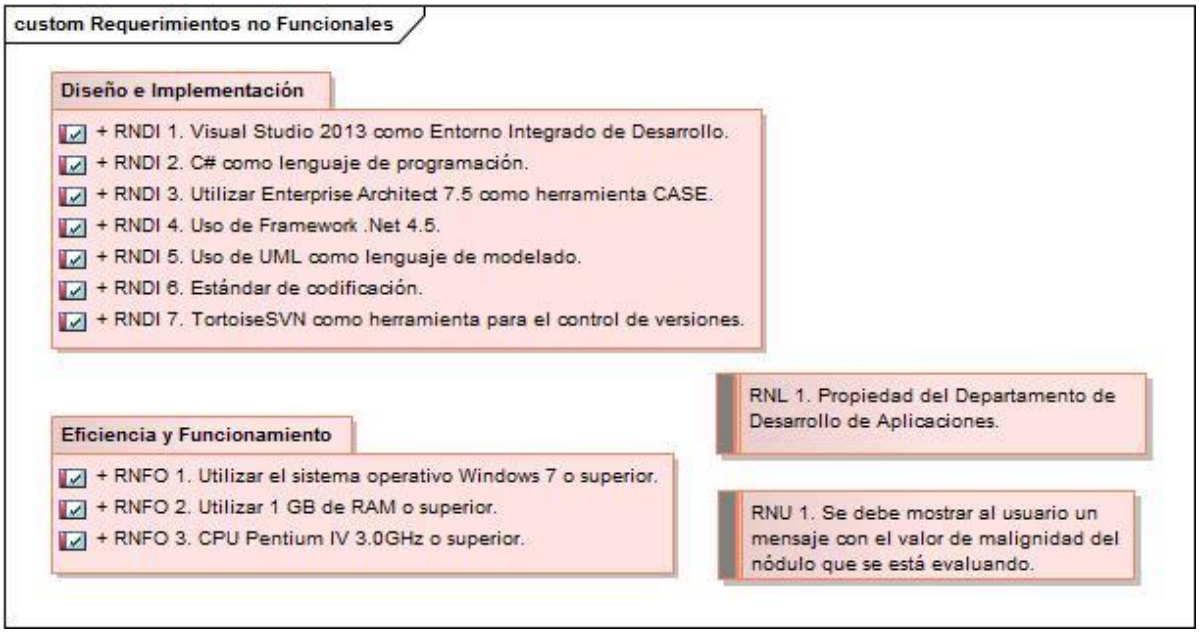


Figura 6. *Requisitos no Funcionales del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

2.5. Propuesta de solución

Para logra la implementación exitosa de un algoritmo de clasificación es necesario tener en cuenta un conjunto de elementos que serán tratados a continuación. Estos elementos comprenden la selección del clasificador a partir de las características de los datos, así como el afinamiento de cada una de las características específicas del algoritmo seleccionado. Un elemento importante es el trabajo realizado para minimizar el impacto de las desventajas propias del modelo elegido en la clasificación.

2.5.1. Selección del clasificador a utilizar en el algoritmo de clasificación de nódulos pulmonares solitarios

La selección del clasificador depende de las características de los datos a utilizar: puede ser supervisado, parcialmente supervisado o sin supervisar. Cada uno corresponde a un estado de conocimiento acerca de las clases objetivos; conocimiento total acerca de la pertenencia de los objetos, parcial o total desconocimiento respectivamente (36, 48).

Para el desarrollo de la presente investigación se cuenta con un conjunto de estudios médicos diagnosticados y se conocen de antemano las clases establecidas en criterios de malignidad que varían desde 1 hasta 5. En este grupo de datos las estructuras nodulares ya están identificadas y sus rasgos han sido cuantificados siguiendo el estándar XML aprobado por el *Lung Image Database Consortium* (LIDC) (75) del año 2009, actualizado en 2010. Teniendo esta información los autores deciden emplear el enfoque de clasificación supervisada.

La clasificación supervisada trabaja con dos hipótesis bien definidas (36):

a) las clases son de naturaleza determinística pues se cuenta con un vector que representa a todos los objetos de una clase y se conoce como vector prototipo.
b) toda la información necesaria y suficiente para su diseño se encuentra disponible de antemano.

2.5.2. Características de la solución

Una vez realizados los estudios pertinentes a los procedimientos desarrollados por varios autores para la realización de sistemas CAD, útiles para la clasificación de estructuras nodulares, se propone la realización de un algoritmo capaz de clasificar nódulos pulmonares solitarios. El algoritmo propuesto procesa la información de un nódulo previamente detectado, en busca de patrones para su clasificación, mostrando el resultado de forma automatizada, devolviendo un valor de malignidad entre 1 y 5. Para obtener la solución deseada se procedió a diseñar el algoritmo de clasificación kNN. Se describen brevemente cada uno de los pasos que el algoritmo ejecuta y se definen los parámetros que serán tenidos en cuenta en su diseño.

El algoritmo kNN inicia su análisis a partir de un conjunto de datos de entrenamiento o instancias ya clasificadas. Al llegar una nueva instancia a ser clasificada, se emplea la función de evaluación que asignará la nueva instancia a una de las clases ya establecidas. Para determinar el valor de la función de evaluación, se calcula la distancia de la nueva instancia respecto a cada una de las instancias con las que cuenta el algoritmo en la base de instancias. A partir del valor establecido para k, se seleccionan las k instancias o k vecinos más cercanos que minimicen la distancia. La clase resultante de la nueva instancia será la que más veces se repita en estos k vecinos más cercanos. En caso de igualdad se debe aplicar una regla de desempate, que finalmente determine la pertenencia a una de las clases previamente definidas.

Las instancias empleadas por el algoritmo están descritas por un vector de características o rasgos que toman valores discretos (76, 77). Para el cálculo de la distancia se pueden emplear medidas de similitud como (78): *euclidean*, *cityblock*, *cosine*, *correlation* y *hamming*. En (78) se desarrolla un experimento empleando validación cruzada para determinar cuál de estas fórmulas arroja los resultados más precisos en la clasificación. En el experimento, la distancia euclidiana no ponderada ($d(x_i,x_j)$) (76) para valores de k que varían desde k=1 hasta k=9, obtiene los mejores resultados. Los autores de la investigación asumen los resultados del experimento y deciden utilizar para la modelación del algoritmo la distancia euclidiana no ponderada.

$$d(x_i,x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2} \qquad [1]$$

La función de evaluación u objetivo (76, 79) está definida como $f(x_q)$ dónde x_q es la nueva instancia a ser clasificada. La misma devuelve valores entre 1 y 5 en representación de cada uno de los niveles de malignidad con que están clasificados los nódulos pulmonares solitarios. El algoritmo kNN solo calcula el valor de la función objetivo general ante la llegada de una nueva instancia para su clasificación.

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(\mathrm{v}, \mathrm{f}(x_i)) \qquad [2]$$

donde δ(a, b)=1 si a=b y δ(a, b)=0 en otro caso.

Para mejorar los resultados obtenidos en la clasificación, varios autores han utilizado diferentes métodos (76, 80, 81). Uno de los más utilizados, es asignarle un peso a la contribución de cada una de las instancias de la clasificación, asignando un mayor peso a las más cercanas a la instancia a clasificar. Para tratar esta situación existen diversas soluciones, las más utilizadas son la distancia inversa (80) y la similaridad (81). Teniendo en cuenta que la función objetivo que se está utilizando es discreta (79), ambas pueden ser empleadas. En un experimento realizado en (79), la distancia inversa obtuvo mejores resultados para la clasificación. A partir de esos resultados obtenidos, en esta investigación el peso del voto de cada vecino se medirá de acuerdo a la distancia inversa respecto a x_q (76, 80), por lo que la función objetivo finalmente quedaría de la siguiente forma.

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} \omega_i \delta(\mathrm{v}, \mathrm{f}(x_i)) \qquad [3]$$

donde

$$\omega_i = \frac{1}{d(x_i x_j)} \qquad [4]$$

En caso de que la instancia a clasificar x_q coincida exactamente con una de las instancias de entrenamiento x_i y la distancia se haga cero, se asignará el valor de $f(x_q) = f(x_i)$ (76, 80).

Teniendo en cuenta que el problema a resolver comprende la clasificación de las estructuras nodulares en cinco clases diferentes, no se puede garantizar un desempate únicamente utilizando un número impar para k (78). De esta manera se debe emplear algoritmos de desempate (82) tales como: *nearest*, *random* o *consensus*. De estos algoritmos en (82) se realizan diferentes pruebas, obteniendo mejores resultados *nearest*.

Los autores de la investigación considerando los resultados de estas pruebas y analizando el peso que tiene en la clasificación el vecino más cercano a la instancia a clasificar, se deciden a emplear el algoritmo *nearest*. Para determinar el número k de vecinos óptimo se tuvo en cuenta el resultado obtenido en (79), en este, para la fórmula objetivo que emplea la distancia inversa como peso el valor de k ideal quedo establecido en 5. Además se desarrolló un experimento en ambiente Matlab con valores de k desde k=3 hasta k=13 y los resultados arrojados aparecen en la Tabla 7:

Tabla 7. *Resultados del experimento para determinar el valor de k que mejores resultados de precisión y sensibilidad arroja (elaborada por los autores).*

Valor del parámetro k	Sensibilidad promedio con peso inverso	Precisión promedio
3	0.94444	0.87142
5	0.92	0.88572
7	0.82436	0.81428
9	0.7819	0.8349
11	0.81006	0.80358
13	0.77214	0.7619

Los resultados relacionados en la Tabla 7 evidencian que la mayor sensibilidad promedio se alcanza para un valor de k =3. El cálculo del indicador sensibilidad por sí solo, no da una medida de cuán preciso fue el proceso diagnóstico. Este elemento fue detectado en el estudio de tendencias, donde varios autores obtienen altos niveles de sensibilidad sin llegar a analizar el número de falsos positivos y falsos negativos encontrados, lo que trae consigo que no se alcancen siempre altos niveles de precisión durante el proceso de clasificación.

Un número elevado de falsos positivos o negativos, implica afectaciones a la calidad de vida de los pacientes erróneamente diagnosticados. Por tal motivo, se decide determinar el valor de k, a partir del análisis de la precisión promedio alcanzada por el algoritmo en las pruebas experimentales realizadas. El mayor valor de precisión promedio fue alcanzado para k=5. Los resultados fueron comparados con los obtenidos en un experimento similar realizado en (78, 79) y los autores de esta investigación arriban a la conclusión de que el valor óptimo de k para el problema planteado es k=5. Para este valor de k, el algoritmo kNN obtiene altos valores de sensibilidad y precisión.

A partir de los elementos descritos anteriormente se puede resumir la estrategia de funcionamiento del algoritmo para la clasificación de nódulos pulmonares solitarios, quedando de la siguiente forma:

Algoritmo de entrenamiento:

- Por cada muestra de ejemplo de la forma {x, f(x)}, adicionarlas a la lista muestras_entrenamiento

Algoritmo de clasificación:

- Dada una nueva instancia x_q a ser clasificada:
  - Calcular la distancia de x_q respecto las muestras de entrenamiento desde x_1 hasta x_n.
  - Seleccionar las k instancias más cercanas a x_q
  - Aplicar la función objetivo

2.5.3. Reducción de instancias mal clasificadas

Una de las características necesarias para la correcta ejecución del algoritmo de clasificación de estructuras nodulares es la limpieza de los datos utilizados en el entrenamiento. Pese a la fortaleza ante datos ruidosos de kNN, se hace necesario depurar la información que será utilizada para su

entrenamiento. Este preprocesamiento de los datos comprende la detección de aquellas instancias que estén mal clasificadas en el conjunto de datos inicial. Diversos autores han propuesto varias estrategias alcanzando distintos resultados (83, 84). Algunos proponen la ejecución de algoritmos de optimización bioinspirados, como *Particle Swarm Optimization* (PSO) (84) y otros proponen el empleo de algoritmos de *clustering;* entre ellos *k Means* ha sido extensivamente utilizado en conjunto con kNN para resolver este problema (85).

Los autores una vez analizadas las diferentes técnicas, deciden emplear el algoritmo *k Means* para realizar el proceso de reducción de falsos positivos, detectando y eliminando las instancias mal clasificadas de la base de instancias. Primeramente se utilizaron las implementaciones de *k Means* que traen las aplicaciones Weka y Matlab. A continuación se compararon los valores obtenidos con los alcanzados en la implementación del algoritmo *k Means* realizada por los autores de la presente investigación. Los resultados arrojados por la comparación muestran una ligera diferencia entre las implementaciones de Matlab y Weka, los resultados obtenidos por la implementación de los autores, mostró mejores resultados al ser capaz de permitir eliminar un mayor número de instancias incorrectamente clasificadas. Pese a que el tiempo de ejecución fue muy superior, los resultados alcanzados permitieron obtener una base de instancias lista para ser utilizada en el entrenamiento del algoritmo.

2.5.4. Determinar rasgos relevantes para el proceso de clasificación de estructuras nodulares

Es importante destacar que el empleo de kNN como algoritmo de clasificación conlleva el riesgo de caer en el error de dimensionalidad. Esto ocurre debido a que, para calcular las distancias emplea todos los atributos de una instancia. Puede darse el caso de que las instancias posean numerosos atributos, pero solo algunos de estos sean relevantes para la clasificación. Esta situación provoca que si las instancias tienen atributos no relevantes iguales en valor, pueden estar distantes en el espacio *n* dimensional, aunque el algoritmo devuelva su cercanía aparente. La distancia entre vecinos estará dominada en este caso por el mayor número de atributos irrelevantes.

Para mitigar el problema de la dimensionalidad se aplican dos estrategias, una básica y otra especializada. En ambas estrategias se asignan pesos a los atributos en dependencia de su impacto en la clasificación:

- La primera comprende la reducción de atributos mediante el empleo de algoritmos que evalúen el impacto en la clasificación de la presencia o ausencia de un atributo (85). La escala empleada es discreta con dos posibles valores: relevante (1) o no relevante (0).
- La segunda estrategia consiste en asignarle pesos o contribuciones a los atributos a la hora de calcular la medida de similitud. El valor exacto de los pesos puede ser determinado por un enfoque de validación cruzada (76, 78) o por la aplicación al conjunto de atributos, de métodos de evaluación tales como (86): *Information Gain, Gain Ratio, Significance Feature Evaluator* y *Relief*. Entre los algoritmos más utilizados para la asignación de pesos a los atributos se encuentran: algoritmos genéticos (GA, por sus siglas en inglés) y *Correlation feature based selection* (CFS). Ambos algoritmos emplean una escala no discreta entre 0 y 1 para evaluar el impacto en la clasificación de los atributos.

Los autores de la investigación deciden adoptar la segunda estrategia. Esta es capaz de evaluar con mayor precisión el grado de relevancia de un atributo en la clasificación, permitiendo alcanzar altos valores de efectividad.

Se realizó un experimento para determinar el subconjunto de características relevantes en el proceso de clasificación. El experimento fue realizado utilizando la aplicación Weka. Teniendo en cuenta las características del juego de datos, se selecciona el evaluador de atributos *CfsSubsetEval* (implementación del algoritmo CFS) (87). Este permite evaluar el impacto de un conjunto de atributos, considerando la habilidad predictiva individual de cada uno y el grado de redundancia entre ellos. Una vez seleccionado el evaluador se probaron los algoritmos *BestFirst* (88) y *GreedyStepwise* (89) para recorrer el espacio de búsqueda. Los atributos seleccionados por los algoritmos y sus pesos correspondientes son los siguientes:

- *BestFirst*: sutileza=1, estructura interna=0.1, calcificación=1, esfericidad=0.2, borde=0.1, lobulación=0, espiculación=1 y textura=0.
- *GreedyStepwise*: sutileza=1, estructura interna=0.1, calcificación=1, esfericidad=0.2, borde=0.1, lobulación=0, espiculación=1 y textura=0.

Una vez analizados los resultados y realizada una intersección de los conjuntos finales se evidencia una coincidencia en la selección de rasgos llevada a cabo por los algoritmos. Los autores de la

investigación deciden no tener en cuenta los atributos (lobulación y textura), estos solo introducen ruido, pues por si solos no tienen impacto en la clasificación.

2.5.5. Funcionamiento general del algoritmo de clasificación de nódulos pulmonares solitarios

Para obtener la solución deseada, la instancia que se desea clasificar es comparada con las instancias almacenadas en la base de conocimientos y la distancia respectiva entre ellas es calculada, Figura 7(a). Para realizar este paso se emplea la distancia euclidiana no ponderada. Una vez obtenidas las distancias se seleccionan las cinco menores distancias, Figura 7(b) y se procede a aplicar la función objetivo.

La función objetivo analiza los valores obtenidos entre la comparación del vector de clases con las clases de cada una de las k instancias seleccionadas. Se obtiene un nuevo vector que representa las cantidades de cada clase que fueron encontradas. Con dicho vector se calcula el argumento máximo o clase con mayor cantidad de repeticiones y ese es el resultado de la clasificación, Figura 7(c). En caso de que al calcular el máximo existan varias clases con igual cantidad se aplica una estrategia de desempate basada en el más cercano, que finalmente determina la clase.

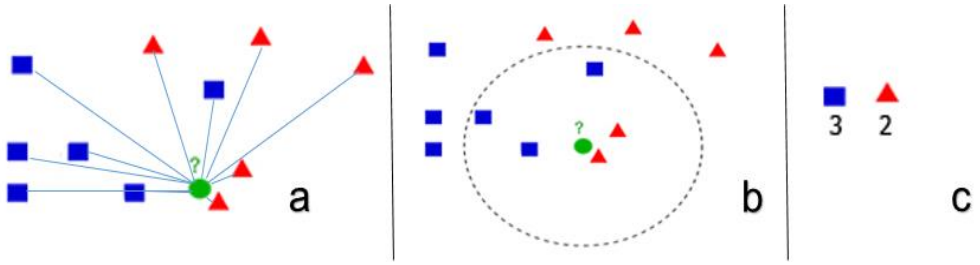


Figura 7. *Proceso de clasificación de nódulos pulmonares solitarios: (a) imagen que representa el cálculo de la distancia desde la instancia a clasificar hasta todas las demás, (b) selección de las cinco instancias más cercanas, (c) vector obtenido con las cantidades respectivas a cada clase (elaborada por los autores).*

2.6. Definición de los actores del Sistema del algoritmo para la clasificación de nódulos pulmonares solitarios

Un actor representa un rol con un comportamiento determinado, estos pueden ser primarios o secundarios. Los actores no solo son personas, sino también, organizaciones, software y computadoras (90). Los actores asociados a la presente investigación se muestran en la Tabla 8.

Tabla 8. *Definición del actor referente al algoritmo de clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

Actor	Descripción
Especialista	Radiólogo que va a realizar la clasificación a partir de los rasgos o descriptores de cada una de las estructuras nodulares.
Algoritmo de detección	El algoritmo que identifica en las imágenes médicas de TC las estructuras nodulares con sus descriptores y se las envía al algoritmo de clasificación.
Usuario General	Es el actor que clasifica las estructuras nodulares.

2.7. Diagrama de Casos de Uso de Sistema

En la Figura 8 se muestra el Diagrama de Casos de Uso del Sistema referente al algoritmo para la clasificación de nódulos pulmonares solitarios. Este diagrama representa el comportamiento y la interacción de los usuarios con el algoritmo.

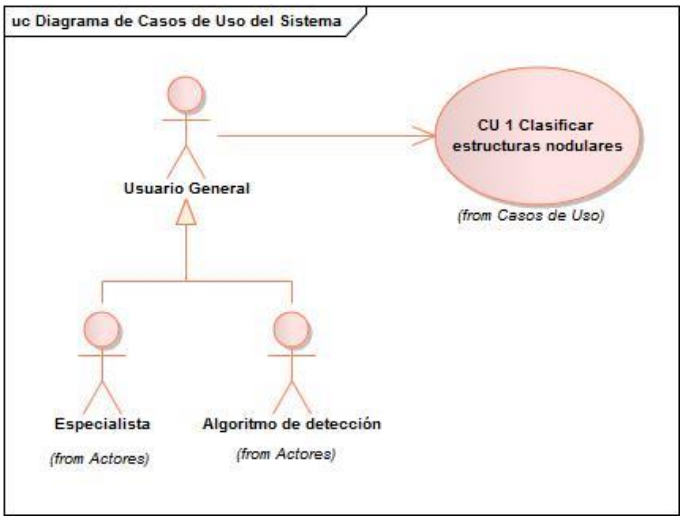


Figura 8. *Diagrama de Casos de Usos del Sistema del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

2.8. Descripción del Caso de Uso del Sistema Clasificar estructuras nodulares

En la Tabla 9 se muestra la descripción del Caso de Uso del Sistema Clasificar estructuras nodulares. La descripción se realiza con el objetivo de especificar cada uno de los elementos que componen el caso de uso así como los flujos de eventos por los que está compuesto. Aparece reflejada la valoración de la complejidad y la prioridad del caso de uso.

Tabla 9. *Descripción del Caso de Uso del Sistema Clasificar estructuras nodulares del algoritmo para la clasificación de estructuras nodulares.*

Objetivo	Devolver el valor de malignidad de los nódulos
Actores	Usuario General (Inicia)
Resumen	El caso de uso comienza cuando el usuario general obtiene las estructuras nodulares con sus descriptores y se los envía al algoritmo de clasificación. El algoritmo devuelve la clasificación del nódulo pulmonar solitario.
Complejidad	Alta
Prioridad	Media
Referencias	RF 1.1, RF 1.2, RF 1.3
Precondiciones	Los descriptores de las estructuras nodulares han sido obtenidos.
Postcondiciones	Se devolvió el valor de la malignidad asociado a la estructura nodular.
Flujo de eventos	
Flujo básico Clasificar estructuras nodulares	
1. El usuario general obtiene una estructura nodular con sus descriptores y ejecuta la opción Clasificar estructuras nodulares. 2. El sistema procesa la información y devuelve la clasificación de la estructura nodular. 3. Termina el caso de uso.	

Flujos alternos		
1ª No se enviaron por parámetros los descriptores		
1. Se intenta ejecutar la opción Clasificar estructuras nodulares sin haber enviado por parámetros una estructura nodular y sus descriptores. *2.* El algoritmo muestra un mensaje de error, "Para clasificar un nódulo inserte el valor de los descriptores correctamente". *3.* Regresar al paso 1 del flujo básico.		
Sección 1: "Calcular distancia euclidiana**"**		
Flujo básico		
1. El sistema calcula la distancia euclidiana entre el nódulo a clasificar y las instancias en la base de conocimientos. 2. Termina el flujo básico.		
Flujos alternos		
No aplica		
Sección 2: "Seleccionar las k instancias más cercanas**"**		
Flujo básico		
1. El sistema selecciona las k instancias más cercanas de la lista de distancias. 2. Termina el flujo básico.		
Flujos alternos		
No aplica		
Sección 2: "Cálculo de la función objetivo**"**		
Flujo básico		
1. El sistema aplica la función objetivo y obtiene el valor de pertenencia del nódulo a su clase. 2. El sistema muestra un mensaje con el valor de la clase del nódulo. 3. Termina el caso de uso.		
Flujos alternos		
No aplica		
Relaciones	**CU incluidos**	-

	CU extendidos	-
Requisitos no funcionales	RNDI 1, RNDI 2, RNDI 4, RNDI 6	
Asuntos pendientes	-	

2.9. Conclusiones del capítulo

Luego de realizada la propuesta de solución para el desarrollo del algoritmo de clasificación de nódulos pulmonares solitarios se llegaron a las siguientes conclusiones:

- A partir del análisis de artículos científicos donde se utiliza el algoritmo kNN y de los resultados del experimento realizado por los autores de la presente investigación utilizando diversos valores de k, se determinó que:
  - El valor óptimo para k es 5.
  - La medida de similitud es la distancia euclidiana no ponderada.
- Por la importancia que tiene para la clasificación la cercanía entre vecinos se decidió utilizar:
  - La función de evaluación modificada con la fórmula de peso inverso.
  - Como algoritmo de desempate el más cercano.
- Se dispuso emplear el algoritmo *k Means* para eliminar las instancias mal clasificadas de la base de instancias.
- Se identificó que los atributos lobulación y textura no son relevantes para el proceso de clasificación.

CAPÍTULO 3. ARQUITECTURA, DISEÑO, IMPLEMENTACIÓN Y VALIDACIÓN DEL ALGORITMO PARA LA CLASIFICACIÓN DE NÓDULOS PULMONARES SOLITARIOS

En este capítulo se plantea el diseño del algoritmo para la clasificación de nódulos pulmonares solitarios, a partir de los Diagramas de Clases del Diseño y el Diagrama de Secuencia del Diseño. Se describen los elementos de estos diagramas, así como el estilo arquitectónico y el patrón de diseño utilizado en el algoritmo. Se muestra el Diagrama de Componentes de la implementación del algoritmo para la clasificación de nódulos pulmonares solitarios. Se reflejan los principales métodos desarrollados para lograr la clasificación de las estructuras nodulares, así como el estándar de codificación utilizado. A partir del diseño de un experimento se presentan los resultados arrojados por el algoritmo, realizando una comparación con los XML previamente diagnosticados disponibles en la base de datos *The Lung Image Database Consortium Image Collection-Image Database Resource Initiative.*

3.1. Diseño del algoritmo de clasificación de nódulos pulmonares solitarios

La etapa de diseño tiene un papel fundamental en el proceso de desarrollo de software. Su principal objetivo es: traducir los requisitos en un diseño que describa el algoritmo para la clasificación de nódulos pulmonares solitarios. Sobre esta etapa recae gran parte de la calidad final del algoritmo. Los diferentes elementos que se manejan en la realización del diseño deben quedar bien especificados para una correcta implementación del algoritmo. (91)

Los Diagramas de Clases del Diseño son parte de los diagramas de estructura. Son utilizados para describir la estructura estática de un sistema porque no describen el comportamiento de este en el tiempo (91). En estos diagramas se modelan la estructura y el comportamiento de las clases, sus atributos y sus asociaciones con otras clases (asociación, generalización, dependencia y realización). Este es el principal diagrama para el análisis y el diseño (92). Mediante el empleo de los Diagramas de Clases del Diseño se puede alcanzar un mejor entendimiento de la estructura del algoritmo para la clasificación de nódulos pulmonares solitarios para su implementación. En la Figura 9 se muestra el Diagrama de Clases del Diseño perteneciente al algoritmo, para una mayor comprensión de las clases del diagrama, se realizó una descripción de las mismas, ver Anexo 1.

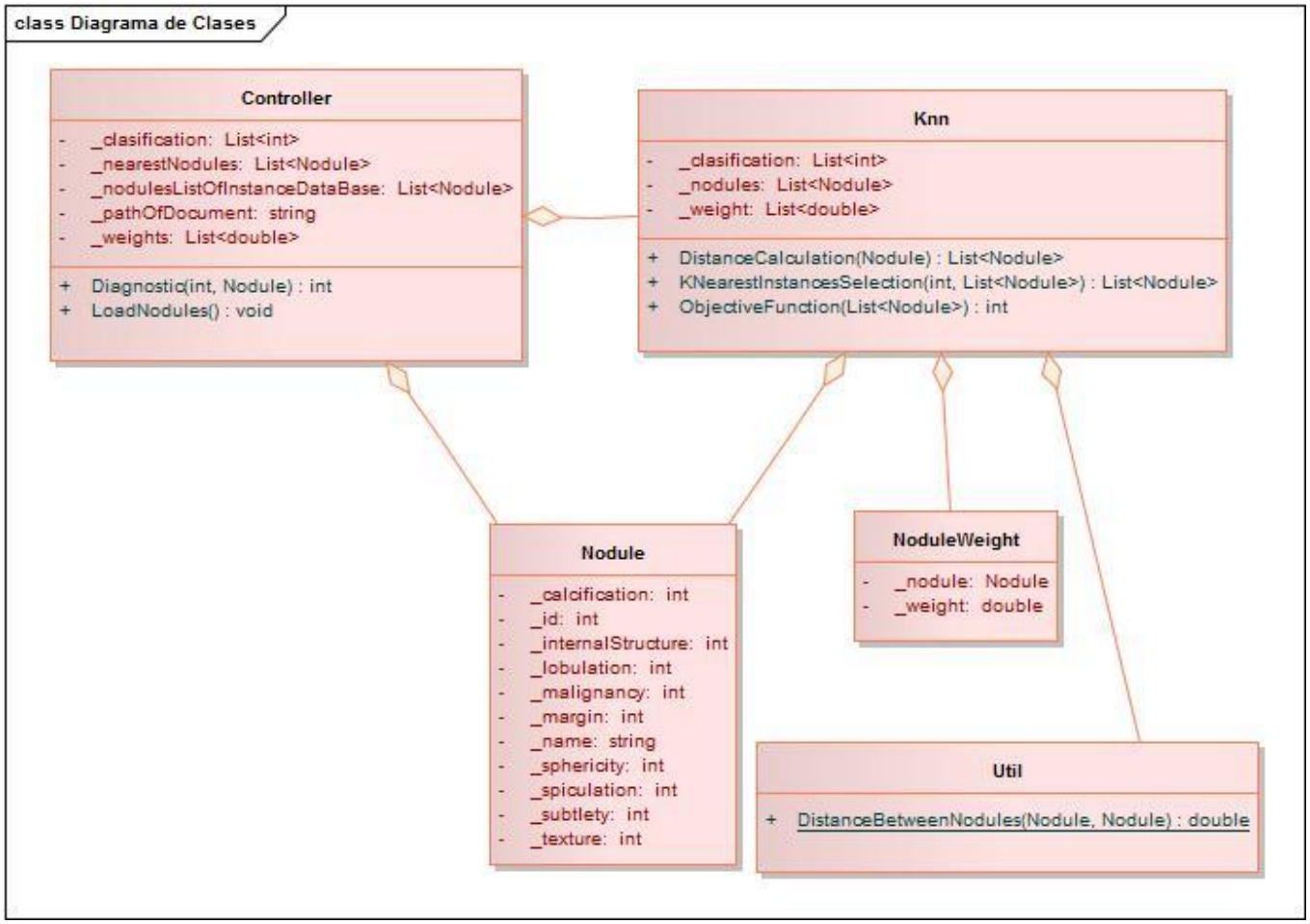


Figura 9. *Diagrama de Clases del Diseño del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores).*

Con el fin de obtener un mayor entendimiento de las actividades que se llevan a cabo en el caso de uso referente al algoritmo para la clasificación de nódulos pulmonares solitarios, se realizó el Diagrama de Secuencia del Diseño del mismo, ver Figura 10. Los Diagramas de Secuencia del Diseño permiten mostrar de forma gráfica las interacciones existentes entre un conjunto de objetos y sus relaciones, incluyendo los mensajes que son enviados entre ellos de forma ordenada en el tiempo (91).

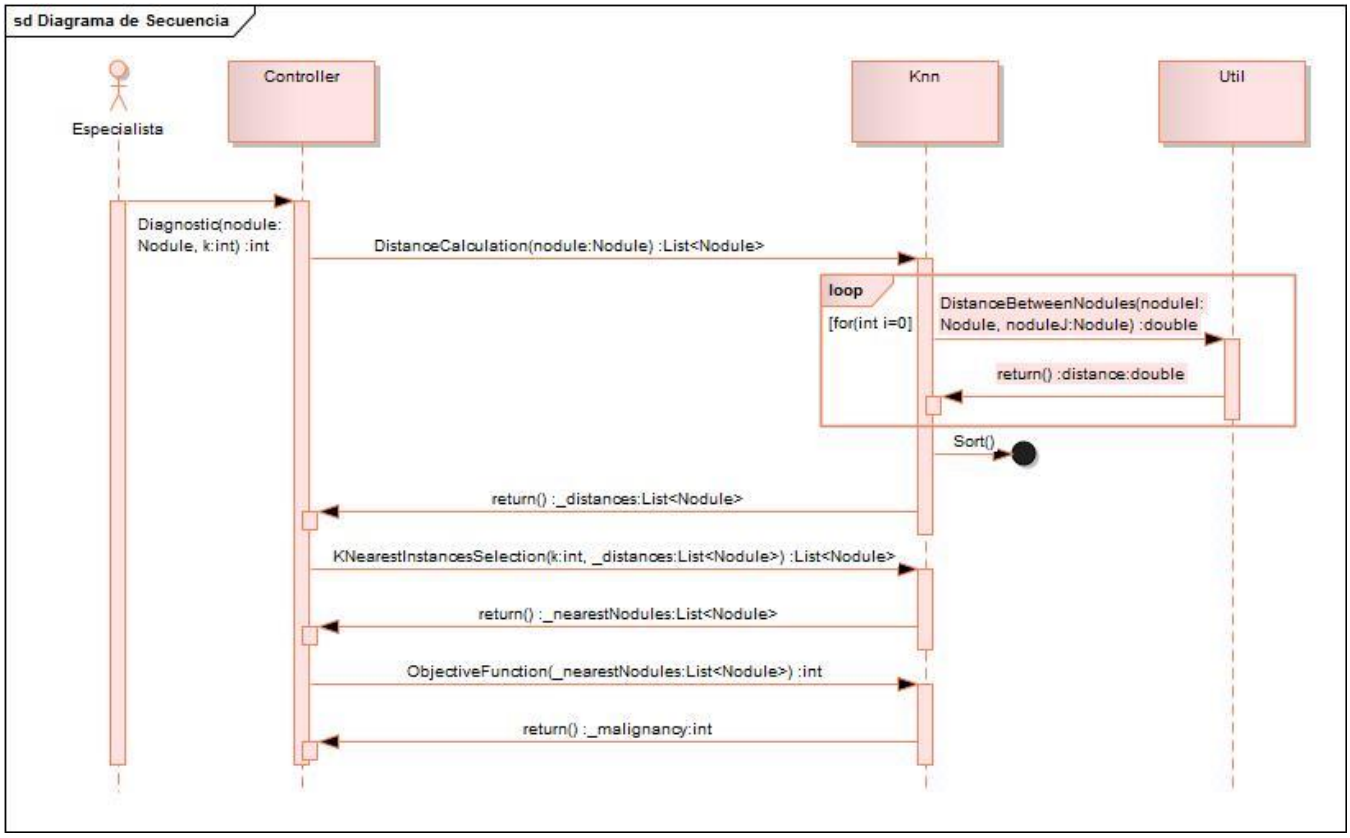


Figura 10. *Diagrama de Secuencia del Diseño del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores).*

Descripción del Diagrama de Secuencia del Diseño del algoritmo para la clasificación de nódulos pulmonares solitarios

Para utilizar el algoritmo para la clasificación de nódulos pulmonares solitarios, el usuario general selecciona la opción “Clasificar nódulos pulmonares” pasándole por parámetros una estructura nodular con su descripción. Para emitir un diagnóstico se realizan un conjunto de pasos que serán descritos a continuación:

1. Se ejecuta el método *Diagnostic* que realiza las llamadas a todos los demás métodos a utilizar. Este método se encuentra en la clase *Controller.*
2. El primer método en ser ejecutado dentro del algoritmo es el *DistanceCalculation* que se encarga de calcular las distancias entre el nódulo a clasificar y los nódulos contenidos en la base de instancias. Este método se implementa en la clase *Knn* y requiere la ejecución del

método *DistanceBetweenNodules* que se encuentra en la clase *Util* y calcula la distancia entre dos nódulos pasados por parámetros. Finalmente se realiza la llamada al método *Sort* para ordenar la lista de distancias de menor a mayor.

3. Una vez obtenida la lista de distancias ordenadas se ejecuta el método *KNearestInstancesSelection*, este se encuentra implementado en la clase *Knn* y se encarga de devolver las k instancias más cercanas al nódulo que se está evaluando.
4. Con las k instancias más cercanas se ejecuta el método *ObjectiveFunction*, el mismo se encuentra implementado en la clase *Knn* y se encarga de evaluar la función objetivo y devolver el valor de malignidad del nódulo que se está evaluando en forma de mensaje.

3.2. Modelo arquitectónico del algoritmo de clasificación de nódulos pulmonares solitarios

La arquitectura de software juega un papel fundamental en la organización de un sistema pues relaciona sus componentes entre ellos y con el entorno (62). Los estilos o modelos arquitectónicos ofrecen soluciones a disímiles problemas de arquitectura de software. Estos dan una descripción de los elementos y el tipo de relación que tienen junto con un conjunto de restricciones sobre cómo pueden ser usados. Expresan un esquema de organización estructural con un nivel de abstracción mayor con respecto a los patrones de diseño (93).

Para el diseño de la arquitectura del algoritmo de clasificación se identificó un estilo arquitectónico dividido por capas. El modelo basado en capas permite estructurar aplicaciones que pueden ser descompuestas en grupos de subtareas, en las que cada grupo de subtareas se encuentra en un nivel específico de abstracción. Cada capa ofrece un conjunto de servicios a las capas superiores e inferiores (94). La necesidad de este modelo surge a partir de que las partes del sistema deben ser intercambiables, lo que significa que los diferentes componentes deben ser capaces de ser reemplazados por implementaciones alternativas si surge la necesidad, sin afectar el resto del sistema. La principal ventaja de este enfoque es que permite incorporar nuevos algoritmos simplemente reescribiendo una delimitada sección del código.

A partir de la ejecución de una llamada al método de clasificación a través de la interfaz del sistema, se origina una petición hacia la capa de procesamiento. En esta se descompone la petición y se obtienen los parámetros que actúan como entrada del proceso de clasificación. La capa de

procesamiento se relaciona directamente con la capa de dominio e interactúa con ella para su funcionamiento. La capa de dominio hace uso de la capa de persistencia para realizar las consultas de obtención de instancias de la base de instancias. Una vez terminado el proceso, el resultado viaja en forma de mensaje hasta alcanzar la capa de interfaz del sistema. Esta última es la encargada a su vez de mostrar el resultado del proceso de clasificación al usuario que hizo la petición. En todo momento el conjunto de capas está relacionado con la plataforma del sistema, ver Figura 11.

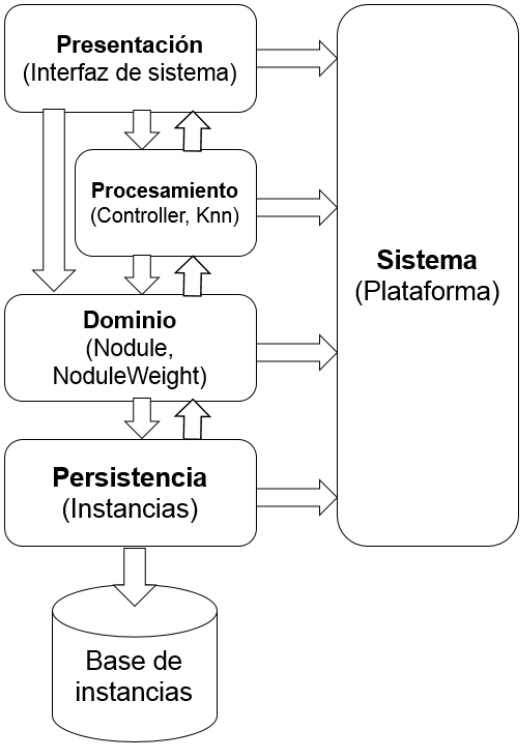


Figura 11. *Modelo arquitectónico basado en capas correspondiente al algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

3.3. Patrón de diseño utilizado en el desarrollo del algoritmo para la clasificación de nódulos pulmonares solitarios

Un patrón de diseño es una buena práctica documentada de la solución de un problema que ha sido aplicado satisfactoriamente en múltiples entornos. Es una solución recurrente a un problema común observado o descubierto durante el estudio o construcción de numerosos software. Su principal

objetivo es incrementar la calidad del software en términos de reusabilidad, mantenimiento y extensibilidad. (62, 91)

En el desarrollo del algoritmo para la clasificación de nódulos pulmonares solitarios se utiliza el patrón Controlador perteneciente a los Patrones de Principios Generales para Asignar Responsabilidades (GRASP, por sus siglas en inglés). El patrón Controlador sirve como intermediario entre una determinada interfaz y el algoritmo que la implementa, recibe los datos del usuario y los envía a las distintas clases según el método que sea llamado. Este patrón sugiere que la lógica de negocios debe estar separada de la capa de presentación para aumentar la reutilización de código y el control sobre la aplicación, así como facilitar las actividades de validación y seguridad. (95) La utilización de este patrón se evidencia en la Figura 12.

Problema: ¿Quién debe ser responsable de manejar la entrada de datos al algoritmo para la clasificación de nódulos pulmonares solitarios?

Solución: usar el patrón Controlador

Aplicación: la clase *Controller* cuenta con el método *Diagnostic*, que ejecuta todos los métodos del proceso de clasificación.

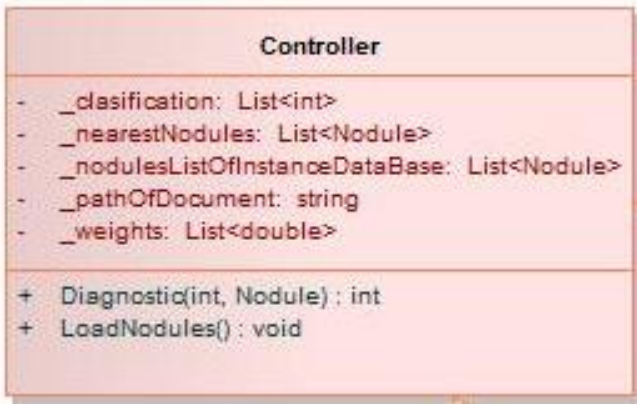


Figura 12. *Utilización del patrón Controlador en el algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

3.4. Diagrama de Componentes del algoritmo de clasificación de nódulos pulmonares solitarios

Un componente representa un módulo físico de código o paquete, puede ser relacionado en un Diagrama de Componentes. Un Diagrama de Componentes muestra varios componentes de un sistema, describiendo sus elementos físicos y sus dependencias (96). En la Figura 13 se muestra el Diagrama de Componentes referente al algoritmo de clasificación de nódulos pulmonares solitarios.

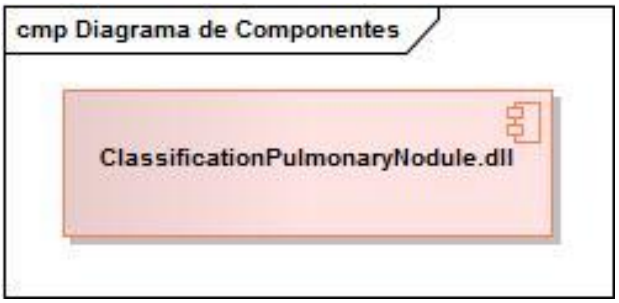


Figura 13. *Diagrama de Componentes del algoritmo para la clasificación de estructuras nodulares (elaborada por los autores)*

Una vez realizado el Diagrama de Componentes se procede a la descripción de cada uno de los elementos con el fin de aumentar la comprensión del mismo, ver Tabla 10.

Tabla 10. *Descripción de los elementos que conforman el Diagrama de Componentes del algoritmo para la clasificación de nódulos pulmonares solitarios (elaborada por los autores)*

Componente	Descripción
ClassificacionPulmonaryNodule.dll	Resultado de la presente investigación. Librería para la clasificación de nódulos pulmonares solitarios en benignos y malignos.

3.5. Estándar de codificación utilizado

Con el objetivo de facilitar a otros desarrolladores el entendimiento del código perteneciente al algoritmo para la clasificación de nódulos pulmonares solitarios y alcanzar una uniformidad en el mismo, se empleó el estándar de codificación que se muestra a continuación:

- Usar nombres descriptivos en inglés para las clases, propiedades y métodos.

- Iniciar el nombre de los atributos de las clases con underscore (_) y letra inicial minúscula, en caso de ser un nombre compuesto se utiliza minúscula y mayúscula. Ejemplos: _nodules, _noduleWeigth.
- Los nombres de los métodos iniciarán con mayúscula y si son palabras compuestas, se empleará notación camel case. Ejemplos: DistanceBetweenNodules, ObjectiveFunction.
- Comentariar las clases y métodos de difícil comprensión.

3.6. Pseudocódigo

Con el objetivo de independizar los algoritmos de un lenguaje de programación en específico y para favorecer su entendimiento se utiliza el pseudocódigo. Su empleo permite la omisión de detalles que no son esenciales para el entendimiento desde el punto de vista humano del algoritmo (97). A continuación se presenta el pseudocódigo del método más significativo del algoritmo de clasificación de estructuras nodulares.

3.6.1. Cálculo de la distancia

El siguiente pseudocódigo se emplea en el cálculo de la distancia entre el nódulo a clasificar y los nódulos de la base de instancias. Ver Figura 14.

```
1  Function DistanceCalculation(Nodule nodule)
2  Inicio
3      byDistanceOrderedList ← emptyList //Creates an empty list
4      mientras counter < instanceListCount hacer
5          temp ← DistanceBetweenNodules(nodule, instancesList[counter])
6   1      byDistanceOrderedList.Adicionar(instancesList[counter],temp)
7          counter ← counter + 1
8      fin mientras
9   2  byDistanceOrderedList.Ordenar() //Orders the list of distances
10     toReturn ← emptyList
11     para cada entry (Nodule, double) in byDistanceOrderedList hacer
12  3      toReturn.Adicionar(entry.Nodule,entry.Distancia,entry.Nodule.Malignancy)
13     fin para
14     alternativeReturn ← toReturn[0]
15  4  si byDistanceOrderedList[0].Distancia = 0 entonces
16         retornar alternativeReturn //Returns first nodule if  distance is 0
17     fin si
18     retornar toReturn
19 Fin
```

Figura 14. *Pseudocódigo del método Calcular Distancia (elaborada por los autores)*

Descripción

1. Se calcula la distancia entre cada nódulo en la base de instancias y el nódulo a clasificar.
2. Se ordenan las distancias de menor a mayor.
3. Se asignan los valores de la distancia a cada nódulo así como su malignidad
4. Se realiza la validación de la distancia 0, si esto se cumple se devuelve la clasificación inmediatamente.

3.7. Validación del algoritmo de clasificación de nódulos pulmonares solitarios.

Para verificar el adecuado funcionamiento del algoritmo de clasificación de nódulos pulmonares solitarios, se hace necesario el cálculo de los indicadores para evaluar el rendimiento del proceso de diagnóstico.

3.7.1. Indicadores para evaluar el rendimiento del proceso de diagnóstico

El resultado de una prueba diagnóstica puede ser positivo o negativo, pero dicho dictamen puede ser correcto o incorrecto, dando lugar a cuatro tipos de resultados (98). Ver Tabla 11.

Tabla 11. *Relación de los indicadores en el estándar de oro o método de referencia (98)*

	Nódulo maligno	Nódulo benigno
Resultado positivo	VP	FP
Resultado negativo	FN	VN

- verdadero positivo (VP): el resultado de la clasificación es positivo en presencia de una anormalidad clínica.
- verdadero negativo (VN): el resultado de la clasificación es negativo en ausencia de la anormalidad clínica.
- falso positivo (FP): el resultado de la clasificación es positivo en ausencia de la anormalidad clínica.
- falso negativo (FN): el resultado de la clasificación es negativo en presencia de la anormalidad clínica.

Para medir el rendimiento del clasificador los elementos anteriores se relacionan en los indicadores sensibilidad (S), especificidad (E) y precisión (P) (98, 99). La sensibilidad y la especificidad son útiles

para medir la efectividad del clasificador (100). La sensibilidad de una prueba diagnóstica es su habilidad para detectar individuos con cáncer. La especificidad es su habilidad para identificar aquellos individuos que no tienen cáncer. La precisión relaciona los indicadores anteriores y brinda una idea general de la eficacia del algoritmo de clasificación.

$$S = VP/ (VP+FN) *100\% \quad [5]$$

$$E = VN/ (VN+FP) *100\% \quad [6]$$

$$P = (VP+VN)/ (VP+VN+FP+FN)*100\ \% \quad [7]$$

En (101) se define un conjunto de etiquetas asociadas a distintos intervalos de precisión como se muestra en la Tabla 12.

Tabla 12. *Relación entre los valores de precisión y los conjuntos que definen (101)*

Precisión	Etiqueta
90%-100%	Excelente
80%-90%	Muy bien
70%-80%	Bien
60%-70%	Suficiente
50%-60%	Malo
< 50%	No es de utilidad

A nivel internacional los valores de precisión comprendidos entre el 80% y el 100% son considerados altos valores de precisión del proceso de diagnóstico.

3.7.2. Análisis del rendimiento clasificatorio del algoritmo de clasificación de estructuras nodulares.

Con el objetivo de analizar los valores de precisión del algoritmo de clasificación de nódulos pulmonares solitarios se realizó un experimento. La población está compuesta por los nódulos que aparecen en las imágenes de la *The Lung Image Database Consortium Image Collection (LIDC/IDRI)*. La *LIDC/IDRI* contiene un total de 1018 casos o estudios realizados. Cada estudio incluye imágenes de Tomografía Computarizada de Tórax y un XML asociado, que guarda los resultados de un proceso de anotación en dos fases, desarrollado por 4 radiólogos torácicos de

experiencia. Las lesiones encontradas se agrupan en 3 clases (nódulo < 3mm, nódulo >= 3mm y no nódulo).

La base de datos contiene 7371 lesiones anotadas como nódulos por al menos un radiólogo. De estas 2669 fueron marcadas como nódulo >= 3mm por al menos un radiólogo, de las cuales 928 (34.7%) recibieron esa marca por los 4 radiólogos. De estos últimos 928 nódulos, realizando una revisión de los XML se identificaron 703 nódulos de los cuales se seleccionó de forma aleatoria un conjunto balanceado de 75 nódulos por cada una de las 5 probabilidades de malignidad para totalizar 375 nódulos a emplear en el experimento. El muestreo es una herramienta de la investigación que permite seleccionar un subgrupo de la población de interés (27). Los elementos en el muestro accidental son seleccionados de forma arbitraria sin tener en cuenta un criterio especial, hasta que se alcance la cantidad deseada.

La ejecución sobre este conjunto de datos del algoritmo *k Means* permitió la reducción del *dataset* a 246 muestras. Quedando la distribución de malignidad de la siguiente forma, ver Tabla 13.

Tabla 13. *Resultados obtenidos al aplicar el algoritmo k Means (elaborada por los autores).*

Clase	Cantidad
1	39
2	56
3	81
4	39
5	31

De las 457 muestras restantes se detectaron 15 muestras cuyos resultados por el algoritmo son no concluyentes al aparecer con igual distancia en clases diferentes. Al ser eliminadas se cuenta con un grupo de pruebas potencial de 442 nódulos. De este conjunto se selecciona arbitrariamente un 15% para pruebas y otro 15% para validación, siendo 65 nódulos para cada uno. Los resultados arrojados aparecen relacionados en las Tablas 14 y 15.

Tabla 14. *Resultados del experimento utilizando el conjunto de pruebas (elaborada por los autores)*

	Maligno	Benigno
Positivo	32	6

	Maligno	Benigno
Negativo	11	16

Tabla 15. *Resultados del experimento utilizando el conjunto de validación (elaborada por los autores)*

	Maligno	Benigno
Positivo	26	12
Negativo	9	18

Del total de nódulos en el conjunto de pruebas fueron correctamente clasificados el 74 %, arrojando el algoritmo valor de sensibilidad, especificidad y precisión de 74%, 73% y 74% respectivamente. Del total de nódulos en el conjunto de validación los valores de sensibilidad, especificidad y precisión fueron de 74%, 60% y 68% respectivamente. Los resultados obtenidos no son altos por el elevado número de FP y de FN obtenidos por lo que es necesario realizar un tratamiento a los datos para mejorar estos indicadores.

3.7.3. Reducción de falsos positivos y falsos negativos.

El número de falsos negativos es evaluado mediante el empleo de una prueba llamada *leave-one-sample-out cross validation (LOSOCV)* (102), en esta prueba cada muestra en el conjunto de entrenamiento es separada y utilizada para probar el rendimiento del clasificador entrenado con las muestras restantes. Por el contrario para evaluar el número de falsos positivos, se utiliza el concepto de prueba nula. En el desarrollo de la prueba mencionada se le presenta al clasificador una muestra que no pertenece a ninguna de las categorías en el conjunto de entrenamiento. Para estas muestras el algoritmo debería devolver una predicción nula debido a que no debería ser asignada a ninguna de las clases conocidas por el clasificador. De lo contrario un error de tipo falso positivo se registra.

Sin embargo, en esta investigación no es posible realizar una prueba nula, debido a la falta de muestras que no pertenecen a ninguna de las categorías en el conjunto de entrenamiento. Un procedimiento alternativo se emplea en estos casos, este se llama *leave-one-class-out cross validation (LOCOCV)* (102). En este enfoque una clase entera es retirada del conjunto de entrenamiento, el clasificador es entrenado con el conjunto restante. Posteriormente el clasificador es probado para comprobar los falsos positivos utilizando las muestras removidas para comprobar.

Esta última prueba solo es aplicable a conjuntos de gran tamaño, pues el eliminar una clase puede influenciar significativamente el rendimiento del clasificador.

Una vez aplicadas ambas estrategias se obtuvo una base de instancias depurada de 182 nódulos y se realizó nuevamente el experimento, obteniéndose los resultados relacionados en la Tabla 16 para 20 iteraciones.

Tabla 16. *Resultados del experimento utilizando el conjunto de pruebas y validación aplicando las estrategias LOSOCV y LOCOCV (elaborada por los autores)*

	Sensibilidad	Especifidad	Precisión	FP	FN
	83,72	77,27	81,53	7	5
	88,88	68,96	80	4	9
	90,62	69,69	80	3	10
	85,29	83,87	84,61	5	5
	85,36	87,5	86,15	6	3
	93,18	57,14	81,53	3	9
	84,84	78,12	81,53	5	7
	89,74	80,76	86,15	4	5
	81,81	76,19	80	8	5
	82,5	80	81,53	7	5
	77,5	88	81,53	9	3
	80	80	80	8	5
	83,78	75	80	6	7
	85,71	73,33	80	5	8
	85,4	70,58	81,53	7	5
	82,85	80	81	6	6
	80,48	79,16	80	8	5
	83,33	73,91	80	7	6
	80	84	81,53	8	4
	83,78	85,71	84,61	6	4
Promedio	84,44	77,46	81,66	6,1	5,8

Los resultados del algoritmo con la base de instancias modificada reducen el número de falsos positivos y falsos negativos respecto a los resultados obtenidos inicialmente. Se alcanza un 81,66% de precisión considerándose un alto valor según lo definido en (101).

3.8. Conclusiones del capítulo

Con la realización de este capítulo se llegaron a las siguientes conclusiones:

- Se definió que el modelo arquitectónico *n* Capas es el más adecuado para la organización del algoritmo de clasificación de nódulos pulmonares solitarios.
- Se determinó el empleo del patrón Controlador para la asignación de responsabilidades.
- La realización del Diagrama de Clases del Diseño facilitó la comprensión del funcionamiento del algoritmo de clasificación.
- El estándar de codificación adoptado permitió la generación de un código claro, de fácil legibilidad y entendimiento por parte de otros desarrolladores.
- Se implementó el algoritmo para la clasificación de nódulos pulmonares solitarios.
- El algoritmo para la clasificación de nódulos pulmonares solitarios arrojó como resultado un 84%, 77% y 81% de sensibilidad, especificidad y precisión respectivamente, al emplear las técnicas LOCOCV y LOSOCV para la reducción del número de falsos positivos y falsos negativos.

CONCLUSIONES

Al término de la investigación los autores arriban a las siguientes conclusiones:

- Se identificó que en el marco de la investigación kNN arroja mejores valores de precisión en el proceso de clasificación que otras técnicas de inteligencia artificial.
- Se dispuso que es necesario eliminar las instancias mal clasificadas de la base de instancias y para ello se empleó el algoritmo *k Means*.
- Se determinó que los atributos lobulación y textura no son relevantes para el proceso de clasificación de nódulos pulmonares solitarios.
- Se desarrolló un algoritmo de clasificación de nódulos pulmonares solitarios que alcanzó los valores de 84%, 77% y 81% de sensibilidad, especificidad y precisión respectivamente.

RECOMENDACIONES

Teniendo en cuenta el estudio realizado durante todo el proceso de desarrollo de la presente investigación y en aras de enriquecer la solución, los autores recomiendan:

- Emplear los algoritmos Optimización por Enjambre de Partículas y Algoritmos Genéticos para la reducción de instancias mal clasificadas y la selección de atributos relevantes respectivamente, para determinar si es factible su adopción en el marco de esta investigación.

REFERENCIAS BIBLIOGRÁFICAS

1. MONTAÑO ZETINA, Luis Manuel. Imagenología y detectores en medicina. *Cinvestav*. 2007. Vol. 2, no. 1, p. 2-3.

2. CHAVARRÍA GÓMEZ, Noelia y PINILLA RAMIRO, David. RESONANCIA MAGNÉTICA. En: Universidad de Alcalá, 2011.

3. OPORTO DÍAZ, Samuel. Detection of Microcalcification in Digital Mammograms by Improved-MMGW Segmentation Algorithm. En: *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*. ISBN 978-0-4799-2235-2.

4. LLOBET AZPITARTE, Rafael, CARLOS PÉREZ CORTÉS, Juan y PAREDES PALACIOS, Roberto. Técnicas Reconocimiento de Formas Aplicadas al Diagnóstico de Cáncer Asistido por Ordenador. *Revista eSalud* [en línea]. 2006. Vol. 2, no. 7. Disponible en: http://archivo.revistaesalud.com/index.php/revistaesalud/article/view/110/260.

5. RAMÍREZ GIRALDO, Juan Carlos, ARBOLEDA CLAVIJO, Carolina y H. MCCOLLOUGH, Cynthia. Tomografía computarizada por rayos X: fundamentos y actualidad. *Revista Ingeniería Biomédica*. 2008. Vol. 2, no. 4.

6. *Universo Médico: modalidades diagnósticas* [en línea]. 2009. Disponible en: http://www.universomedico.com.mx/que-hace-y-que-atiende-cada-especialidad/%C2%BFque-atiende-cada-una-de-las-especialidades/.

7. VALLES BLÁZQUEZ, Beatriz y IZQUIERDO VARELA, Raúl. *DIAGNOSIS DECISION SUPPORT SYSTEMS*. Leganés, España, 2010.

8. HAAR, Liza y ALAVI, Abass. Modalities. *Diagnostic Imaging* [en línea]. 2015. Disponible en: http://www.diagnosticimaging.com/modalities

9. *Cancer Fact Sheet* [en línea]. GLOBOCAN, 2008. Disponible en: http://globocan.iarc.fr/factsheet.asp

10. W. STEWART, Bernard y P. WILD, Cristopher. *Cancer facts sheet* [en línea]. Scientific. World Health Organization, 2015. Disponible en: http://www.who.int/mediacentre/factsheets/fs297/en/

11. CLÚA CALDERÍN, Ana Margarita y GUTIÉRREZ CAMPO, Léster. ANUARIO ESTADÍSTICO DE SALUD. [en línea]. 2015. Disponible en: http://temas.sld.cu/redenfermeriacomunitaria/2015/04/10/anuario-estadistico-de-salud-2014/

12. BOGONI, Luca y P. KO, Jane. Impact of a Computer-Aided Detection (CAD) System Integrated into a Picture Archiving and Communication System (PACS) on Reader Sensitivity and Efficiency for the Detection of Lung Nodules in Thoracic CT Exams. *Journal of Digital Imaging*. 2012. Vol. 25, p. 771-781. DOI 10.1007/s10278-012-9496-0.

13. ALAVI, Asif y MOSENIFAR, Zab. Solitary Pulmonary Nodule. *Medscape*. 2013. Vol. 1, p. 1-4.

14. SUÁREZ CUENCA, Jorge Juan. *Desarrollo de un sistema de diagnóstico asistido por computador para detección de nódulos pulmonares en tomografía computarizada multicorte*. Tesis PhD. 2009.

15. BORGES GONZÁLEZ, Yosuani y NARANJO GORRÍN, Yoanny. *Algoritmo para la identificación de nódulos pulmonares solitarios en imágenes de tomografía de tórax.* Universidad de las Ciencias Informáticas, 2014.

16. RIVERO CASTRO, Arelys. *Algoritmo basado en técnicas de segmentación de imágenes de tomografía de tórax para aumentar el acierto en la identificación de nódulos pulmonares solitarios.* [en línea]. La Habana, Cuba: Universidad de las Ciencias Informáticas, 2014. Disponible en: http://repositorio_institucional.uci.cu/jspui/handle/ident/8550

17. BHUVANESWARI, C., ARUNA, P. y LOGANATHAN, D. A new fusion model for classification of the lung diseases using genetic algorithm. *Egyptian Informatics Journal*. 2014. Vol. 15, no. 2, p. 69–77. DOI 10.1016/j.eij.2014.05.001.

18. NAMIN, S.T., MOGHADDAM, H.A., IZQUIERDO, R. y ESMAEIL-ZADEH, M. Automated detection and classification of pulmonary nodules in 3D thoracic CT images. *Systems Man and Cybernetics*. 2010. Vol. 11, no. 66, p. 3774 - 3779. DOI 10.1109/ICSMC.2010.5641820.

19. CHEN, Hui, WANG, Xiao-hua, MA, Da-qing y MA, Bin-rong. Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pulmonary nodules by computed tomography. *Chinese Medical Journal*. 2007. Vol. 120, no. 14, p. 1211:1215.

20. OUGIAROGLOU, Stefanos, NANOPOULOS, Alexandros, PAPADOPOULOS, Apostolos N. y WELZER-DRUZOVEC, Tatjana. *Adaptative k-Nearest-Neighbor Classification Using a Dynamic Number of Nearest Neighbors*. Scientific. Faculty of Electrical Eng. and Computer Science: University of Maribor, 2011.

21. DENNIS M.C. IDELER. 1: *Comparison of Artificial Neural Network and k-Nearest Neighbor for Classification*. Scientific. Computer Science Department: Brock University, 2010.

22. A. BETANCOURT, Gustavo. LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs). *Scientia et Technica*. 2005. Vol. 11, no. 27.

23. MUSTAFA, Mahfuzah, TAIB NASIR, Mohd, MURAT, Zunairah y SULAIMAN, Norizam. Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image. *Journal of Computer Science & Computational Mathematics*. 2012. Vol. 2, no. 4.

24. SONG, Zhifei y GU, Qi. *Image Classification Using SVM, KNN and Performance Comparison with Logistic Regression* [en línea]. Ciencia. Universidad de Dartmouth, 2011. Disponible en: http://www.cs.dartmouth.edu/~afra/courses/44/w09/project/report/gu-song-report.pdf

25. BEYER, Kevin S., GOLDSTEIN, Jonathan, RAMAKRISHAN, Raghu y SHAFT, Uri. When is a nearest neighbor meaningful? En: *99 Proceedings of the 7th International Conference of Database Theory* [en línea]. Springer-Verlag, 1999. p. 217-235. ISBN 3-540-65452-6. Disponible en: http://dl.acm.org/citation.cfm?id=645503.656271

26. MALDONADO, Sebastián y WEBER, Richard. A wrapper method for feature selection using Support Vector Machines. *Information Sciences*. 2009. Vol. 2, no. 1, p. 2208-2216. DOI 10.1016/j.ins.2009.02.014.

27. HERNÁNDEZ SAMPIERI, Roberto, FERNÁNDEZ-COLLADO, Carlos y BAPTISTA LUCIO, Pilar. *Metodología de la investigación*. 4. México: Mc Graw Hill, 2010. ISBN 970-10-5753-8.

28. CONDE VALERO, A y NAVASCUÉS MARTÍNEZ, E. Estudio del nódulo pulmonar solitario. *Manual de Neumología Neumosur*. 2009. Vol. 3, p. 233-242.

29. QUESADA GONZÁLEZ, Guillermo, OTERO ECHEMENDIA, Yanine y CONDE REBOSO, Anay. Manejo del nódulo pulmonar solitario. *Gaceta Médica Espirituana*. 2010. Vol. 12, no. 2, p. 1-7.

30. SEPÚLVEDA, Cristián, SEPÚLVEDA, Alfredo y FUENTES, Esteban. Nódulo Pulmonar Solitario. *Rev. Chilena de Cirugía.* 2008. Vol. 60, no. 1, p. 71-78.

31. JEONG, Y J, YI, C A y LEE, K S. Solitary Pulmonary Nodules: Detection, Characterization, and Guidance for Further Diagnostic Workup and Treatment. *Radiology*. 2008. Vol. 50, no. 3, p. 2-8. DOI 10.1016/S0033-8338(08)71964-7.

32. SUMMERS, R. M. Road maps for advancement of radiologic computer-aided detection in the 21st century. *Radiology*. 2003. Vol. 229, no. 1, p. 11-13.

33. MASOOMEH, Barzegari. Computer-aided dermoscopy for diagnosis of melanoma. *BMC Dermatology*. 2005. Vol. 5.

34. VAN GINNEKEN, B, SCHAEFER-PROKOP, CM y PROKOP, M. Computer aided diagnosis: how to move from the laboratory to the clinic. *Radiology*. 2011. Vol. 261, no. 3, p. 719-732.

35. ABERLE, DR, ADAMS, AM, BERG, CD, BALCK, WC, CLAPP, JD y FAGERSTROM, RM. Reduced lung-cancer mortality with low-dose computed tomographic screening. National Lung Screening Trial Research Team. *National English Journal of Medicine*. 2011. Vol. 365, p. 395-409.

36. RUIZ SHULCLOPER, José, GUZMÁN ARENAS, Adolfo y MARTÍNEZ TRINIDAD, J. Francisco. *Enfoque Lógico Combinatorio al Reconocimiento de Patrones. Selección de variables y clasificación supervisada.* Colección de Ciencias de la Computación. Instituto Politécnico Nacional de México: Departamento de Ingeniería Eléctrica de México, 1999. ISBN 970-18-2384-1.

37. CHEN, Hui, WANG, Xiao-hua, MA, Da-qing y MA, Bin-rong. Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pulmonary nodules by computed tomography. *Chinese Medical Journal*. 2007. Vol. 120, no. 14, p. 1211:1215.

38. FAN, Jerome, UPADHYE, Suneel y WORSTER, Andrew. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*. 2006. Vol. 8, no. 1, p. 19-20. DOI 10.1177/0956797614541991.

39. MORÉ, Jorge J. The Levenberg-Marquardt algorithm: Implementation and theory. En: WATSON, G. A. (ed.), *Numerical Analysis* [en línea]. Springer Berlin Heidelberg, 1978. p. 105-116. Lecture Notes in Mathematics, 630. ISBN 978-3-540-35972-2. Disponible en: http://link.springer.com/chapter/10.1007/BFb0067700

40. MATSUKI, Yuichi, NAKAMURA, Katsumi, WATANABE, Hideyuki, AOKI, Takatoshi, NAKATA, Hajime, KATSURAGAWA, Shigehiko y DOI, Kunio. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. *AJR. American journal of roentgenology*. 2002. Vol. 178, no. 3, p. 657-663.

41. CAMPADELLI, Paola, CASIRAGHI, Elena y VALENTINI, Giorgio. *Support Vector Machines for Candidate Nodules Classification*. Scientific. Departamento de Ciencias de la Información: Universidad de Milán, Italia, 2005.

42. LIU, L., WANG, K., WEN, D.-W. y LI, Y. Multiple kernel MtLSSVM and its application in lung nodule recognition. *Journal of Jilin University (Engineering and Technology Edition)*. 2014. Vol. 44, no. 2, p. 508-515. DOI 10.13229/j.cnki.jdxbgxb201402037.

43. HAMADA, R. H. Al-Absi, BRAHIM BELHAOUARI, Samir y SUZIAH, Sulaiman. A Computer Aided Diagnosis System for Lung Cancer based on Statistical and Machine Learning Techniques. *Journal of Computers*. 2014. Vol. 9, no. 2, p. 425-431. DOI 10.4304/jcp.9.2.425-431.

44. CHAKRABORTY, Dipanjan. K-Nearest Neighbor Learning. *Computer S 572* [en línea]. Arizona State University. 2014. Disponible en: http://www.csee.umbc.edu/courses/671/fall01/class-notes/k-nn.ppt

45. *StatSoft Electronic Statistics Textbook*. 2010.

46. SYKES, A. O. Regression Analysis. [en línea]. University of Chicago, 2009. Disponible en: http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf

47. MORO, Quiliano y SIMÓN HURTADO, Aránzazu. Introducción al Diseño de Experimentos para el Reconocimiento de Patrones. 2006.

48. BONIFACIO MARTÍN DEL RÍO y ALFREDO SANZ MOLINA. *Redes Neuronales y Sistemas Borrosos*. 2. Zaragoza, España: Alfaomega, 2001.

49. TU, Jack V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*. 1996. Vol. 49, no. 1, p. 1225-1231. DOI 10.1016/S0895-4356(96)00002-9.

50. JORGE MATICH, Damián. *Redes Neuronales: Conceptos Básicos y Aplicaciones.* 2001. Grupo de Investigación Aplicada a la Ingeniería Química (GIAIQ).

51. IGOR V. TETKO, DAVID J. LIVINGSTONE y ALEXANDER I. LUIK. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci*. 1995. Vol. 35, no. 5, p. 826-833. DOI 10.1021/ci00027a006.

52. DAVID H. ALMAN y LIAO NINGFANG. Overtraining in back-propagation neural networks: A CRT color calibration example. *Color Research & Application*. 2002. Vol. 27, no. 2, p. 122-125. DOI 10.1002/col.10027.

53. FABRICE COLAS y PAVEL BRAZDIL. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. *Artificial Intelligence in Theory and Practice*. 2006. Vol. 217, p. 172-177. DOI 10.1007/978-0-387-34747-9_18.

54. MEYER, CR, ARMATO, SG III, FENIMORE, CP, MCLENNAN, G, BIDAUT, LM, BARBORIAK, DP, GAVRIELIDES, MA, JACKSON, EF, MCNITT-GRAY, MF, KINAHAN, PE, PETRICK, N y ZHAO, B. Quantitative imaging to assess tumor response to therapy: common themes of measurement, truth data, and error sources. *Translational Oncology*. 2009. Vol. 2, no. 4, p. 198-210.

55. ARMATO, SG III, MEYER, CR, MCNITT-GRAY, MF, MCLENNAN, G y REEVES, AP. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: A resource for the development of change analysis software. *Clinical pharmacology and terapeutics*. 2008. Vol. 84, no. 4, p. 448-456. DOI 10.1038/clpt.2008.161.

56. KURAVATI, Rohit, SASIDHAR, B y RAMESH BABU, D R. A Novel Method for Classification of Lung Nodules as Benign and Malignant using Artificial Neural Network. *International Journal of Engineering and Computer Science*. 2014. Vol. 3, no. 8, p. 7641-7645.

57. ARMATO, SG III, MC LENNAN, G, BIDAUT, L, MCNITT-GRAY, MF, MEYER, CR, REEVES, AP, ZHAO, B, ABERLE, DR, HENSCKE, CI, HOFFMAN, EA, KAZEROONI, EA, MACMAHON, H, VAN BEEK, EJR y YANKELEVITZ, D. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). A completed reference database of lung nodules on CT scans. *Medical Physics*. 2011. Vol. 38, no. 2, p. 915-931. DOI 10.1118/1.3528204.

58. SHIRAISHI, J, KATSURAGAWA, Shigehiko, IKEZOE, J, MATSUMOTO, T, KOBAYASHI, T, KOMATSU, K, MATSUI, M, FUJITA, H, KODERA, Y y DOI, Kunio. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*. 2000. Vol. 174, no. 1, p. 71-7.

59. CLARK, K, GIERADA, D, MOORE, S y MAFFITT, D. Creation of a CT Image Library for the Lung Screening Study of the National Lung Screening Trial. *Journal of Digital Imaging*. 2010. Vol. 20, no. 1, p. 23-31.

60. CLARK, K, VENDT, B, SMITH, K, FREYMANN, J, KIRBY, J, KOPPEL, P, MOORE, S, PHILLIPS, S, MAFFITT, D, PRINGLE, M, TARBOX, L y PRIOR, F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*. 2013. Vol. 26, no. 6, p. 1045-1057.

61. WEST, David. *Planning a Project with the Rational Unified Process* [en línea]. 2002. Rational Software White Paper. Disponible en: http://www.nyu.edu/classes/jcf/CSCI-GA.2440-001/handouts/PlanningProjWithRUP.pdf

62. JACOBSON, Ivar, BOOCH, Grady y RUMBAUGH, James. *El Proceso Unificado de Desarrollo de Software*. Madrid: Rational Software Corporation, 2000. ISBN 84-7829-036-2.

63. FAVRE, Liliana. *UML and the Unified Process*. Universidad Nacional del Centro de la Provincia de Buenos Aires: IRM Press, 2003. ISBN 1-931777-44-6.

64. EVITTS, Paul. *A UML Pattern Language*. First Edition. New Riders Publishing, 2000. ISBN 1-57870-118-X.

65. Welcome to Visual Studio 2013. [en línea]. 2013. Disponible en: http://msdn.microsoft.com/en-us/library/dd831853.aspx

66. New C# Features in the .NET Framework 4. [en línea]. 2010. Disponible en: http://msdn.microsoft.com/es-es/magazine/ff796223.aspx

67. Enterprise Architect - Herramienta de diseño UML. [en línea]. 2012. Disponible en: http://www.sparxsystems.com.ar/products/ea.html

68. TortoiseSVN. The coolest interface to (Sub)version control. [en línea]. 2014. Disponible en: http://tortoisesvn.net/

69. Matlab. The Language of technical Computing. [en línea]. 2014. Disponible en: http://www.mathworks.com/products/matlab/

70. Weka 3: Data Mining Software in Java. [en línea]. 2013. Disponible en: http://www.cs.waikato.ac.nz/%7Eml/weka/index.html

71. LÓPEZ PÉREZ, Carmelo. Modelo de Madurez de la Capacidad del Software1. *Revista de Ingeniería Informática del CIIRM*. 2004. Vol. 1, no. 1, p. 5-9.

72. ROYCE, Walker. CMM vs. CMMI: From Conventional to Modern Software Management. En: [en línea]. Rational Software Corporation, 2002. p. 3-8. Disponible en: http://www.therationaledge.com/content/feb_02/f_conventionalToModern_wr.html

73. LARMAN, Craig. *UML y Patrones. Una introducción al análisis y diseño orientado a objeto y al proceso unificado* [en línea]. 2. Prentice Hall, 2003. ISBN 978-8420534381. Disponible en: http://www.amazon.es/Uml-y-patrones-Craig-Larman/dp/8420534382

74. SOMMERVILLE, Ian. *Ingeniería del software* [en línea]. 7ma. Pearson Educación S.A., 2009. ISBN 84-7829-074-5. Disponible en: http://zeus.inf.ucv.cl/~bcrawford/Modelado%20UML/Ingenieria%20del%20Software%207ma.%20Ed.%20-%20Ian%20Sommerville.pdf

75. Lung Image Database Consortium (LIDC). *Lung Image Database Consortium (LIDC)* [en línea]. 2014. Disponible en: http://imaging.cancer.gov/programsandresources/informationsystems/lidc

76. MITCHELL, Tom M. *Machine Learning*. New York: McGraw-Hill Science, 1997. ISBN 0070428077.

77. LÓPEZ DÍAZ, José Carlos. *UN ALGORITMO GENÉTICO CON CODIFICACIÓN REAL PARA LA EVOLUCIÓN DE TRANSFORMACIONES LINEALES* [en línea]. Proyecto de Fin de Carrera. Leganés, España, 2010. Disponible en: http://www.bibliopedant.com/OfNRMsISQ1ndjwR6N7Op

78. SAINI, Indu, SINGH, Dilbag y KHOSLA, Arun. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research*. 2013. Vol. 4, no. 4, p. 331-344. DOI 10.1016/j.jare.2012.05.007.

79. BATISTA, Gustavo y FURTADO SILVA, Diego. *How k-Nearest Neighbor Parameters Affect its Performance*. Scientific. Laboratorio de Inteligencia Computacional: Instituto de Ciencias Matemáticas y de Computación, Universidad de Sao Paulo, 2009.

80. WILSON, D.R. y MARTINEZ, T.R. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*. 1997. Vol. 6, p. 1-34.

81. STANFILL, C. y WALTZ, D. Instance-based Learning Algorithms. *Communications of the ACM*. 1986. Vol. 12, p. 1213-1228.

82. R2014 Matlab Documentation. [en línea]. 2014. Disponible en: http://www.mathworks.com/help/bioinfo/ref/knnclassify.html

83. KETTANI, Omar, TADILI, Benaissa y RAMDANI, Faycal. A Deterministic K-means Algorithm based on Nearest Neighbor Search. *International Journal of Computer Applications*. 2013. Vol. 63, no. 15, p. 34-39.

84. SIVAKUMAR, S. y CHANDRASEKAR, C. Modified PSO Based Feature Selection for Classification of Lung CT Images. *International Journal of Computer Science and Information Technologies*. 2014. Vol. 5, no. 4, p. 2095-2098.

85. GOWDA KAREGOWDA, Asha, JAYARAM, M. A. y MANJUNATH, A. S. Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. *International Journal of Engineering and Advanced Technology*. 2012. Vol. 1, no. 3, p. 142-157.

86. GOWDA KAREGOWDA, Asha. Enhancing Performance of KNN Classifier by Means of Genetic Algorithm and Particle Swarm Optimization. *International Journal of Advance Foundation and Research in Computer*. 2014. Vol. 1, no. 5, p. 24-36.

87. HALL, M. A. *Correlation-based Feature Subset Selection for Machine Learning*. New Zealand: Hamilton, 1998.

88. KIRA, Kenji y RENDELL, Larry. A Practical Approach to Feature Selection. En: *Ninth International Workshop on Machine Learning*. 1992.

89. KOHAVI, Ron y JOHN, George. Wrappers for feature subset selection. *Artificial Intelligence.* 1997. Vol. 97, no. 2, p. 273-324.

90. LARMAN, Craig. *Applying UML and Patterns. An introduction to Object-Oriented Analysis and Design and the Unified Process*. 2da. 2002.

91. KUCHANA, Partha. *Software Architecture Design Patterns in Java*. New York: CRC/AUERBACH PUBLICATIONS, 2004. ISBN 0-8493-2142-5.

92. LORENZO CASTILLO, Juan. Diagrama de Clases. *Diseño de software*. Universidad de Jaen. 2006.

93. RAMOS, Juan Carlos y DEPETRIS, Natalia. Estilos Arquitectónicos. *Diseño de Software basado en Arquitecturas*. UTN-FRSF. 2012.

94. BUSCHMANN, Frank, MAUNIER, Regine, ROHNERT, Hans, SOMMERLAD, Peter y STAL, Michael. *Pattern Oriented Software Architecture*. Germany: John Wiley and Sons, 1996. ISBN 0 471 95869 7.

95. SPIEGEL, Jhon. GRASP Design Patterns: Designing Objects with Responsibilities. KutzTown. 2004.

96. BOOCH, Grady, JACOBSON, Ivar y RUMBAUGH, James. *UML Distilled Second Edition. A brief guide to the standard Object Modeling Language*. Addison-Wesley, 2000. ISBN 020165783X.

97. ROY, Geoffrey G. Designing and explaining programs with a literate pseudocode. *Journal on Educational Resources in Computing* [en línea]. 2006. Vol. 6, no. 1. DOI 10.1145/1217862.1217863. Disponible en: http://dl.acm.org/citation.cfm?id=1217862.1217863&coll=DL&dl=GUIDE&CFID=645389442&CFTOKEN=27175044

98. SELVARAJ, H., THAMARAI, S., SELVATHI, D. y GEWALL, L. Brain MRI Slices Classification Using Least Squares Support Vector Machine. *Journal of Theoretical and Applied Information Technology*. 2007. Vol. 1, no. 1, p. 21-33.

99. NITHYA, R. y SANTHI, B. Mammogram Classification Using Maximum Difference Feature Selection Method. *Journal of Theoretical and Applied Information Technology*. 2011. Vol. 33, p. 197-204.

100. ALTMAN, D. G. y BLAND, J. M. Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal*. 1994. Vol. 308, p. 1552-1554.

101. ŠIMUNDIĆ, Ana-Maria. *Measures of diagnostic accuracy: basic definitions* [en línea]. Department of Molecular Diagnostics: University Department of Chemistry, Croatia, 2008. Disponible en: www.ifcc.org/ifccfiles/docs/190404200805.pdf

102. GE, Xijin, TSUTSUMI, Shuichi, ABURATANI, Hiroyuki y IWATA, Shuchi. Reducing False Positives in Molecular Pattern Recognition. *Genome Informatics*. 2003. Vol. 14, p. 34-43.

ANEXOS

1 DESCRIPCIÓN DE LAS CLASES DEL DISEÑO DEL ALGORITMO DE CLASIFICACIÓN DE NÓDULOS PULMONARES SOLITARIOS

Nombre: Controller	
Tipo de clase: Controladora	
Atributo	Tipo
_nodulesListOfInstanceDataBase	List<Nodule>
_nearestNodules	List<Nodule>
_weights	List<double>
_classification	List<int>
_contOfId	int
_pathOfDocument	string
Por cada responsabilidad	
Nombre:	Diagnostic
Descripción:	Método principal del algoritmo, ejecuta las funcionalidades necesarias para calcular la función objetivo que clasificar las estructuras nodulares.

Nombre: KNN	
Tipo de clase: Controladora	
Atributo	Tipo
_nodules	List<Nodule>
_weight	List<double>
_classification	List<int>
Por cada responsabilidad	
Nombre:	DistanceCalculation
Descripción:	Método que calcula distancia entre el nódulo a clasificar y los nódulos de la base de instancias.

Nombre:	KNearestInstancesSelection
Descripción:	Selecciona las k instancias más cercanas al nódulo a clasificar.
Nombre:	ObjectiveFunction
Descripción:	Calcula la malignidad del nódulo basándose en el vector de malignidad.
Nombre:	DistanceBetweenNodules
Descripción:	Calcula la distancia entre dos nódulos cualesquiera.

GLOSARIO DE TÉRMINOS

2D: 2 dimensiones, bidimensional.

3D: 3 dimensiones, tridimensional.

Algoritmo: es un conjunto ordenado y finito de operaciones sistemáticas que permiten el cálculo y hallar la solución a un tipo de problema.

Imagen: representación, semejanza y apariencia de algo. En computación es formada por la unión de MxN píxeles (imagen 2D) o vóxeles (imagen 3D).

Imagenología: comprende la realización de todo tipo de exámenes diagnósticos y terapéuticos en los cuales se utilizan equipos que reproducen imágenes del organismo.

Ministerio de Salud Pública (MINSAP): organismo rector del Sistema Nacional de Salud de Cuba, encargado de dirigir, ejecutar y controlar la aplicación de la política del Estado y del Gobierno en cuanto a la salud pública, el desarrollo de las Ciencias Médicas y la industria médico-farmacéutica.

Patrón arquitectónico: expresa un esquema estructural fundamental de la organización para un sistema de software, que consiste en subsistemas, sus responsabilidades e interrelaciones.

Organización Mundial de la Salud (OMS): es el organismo de la Organización de las Naciones Unidas (ONU) especializado en gestionar políticas de prevención, promoción e intervención en salud a nivel mundial.

Requisitos: capacidades, condiciones o cualidades que la presente investigación debe cumplir y tener.

RNDI: Requisitos no Funcionales de Diseño e Implementación.

RNFO: Requisitos no Funcionales de Funcionamiento.

RNL: Requisitos no Funcionales Legales.

RNU: Requisitos no Funcionales de Usabilidad.

Tácito: expresa un conocimiento implícito del dominio.

XML: Lenguaje de Marcas Ampliable (Extensible Markup Language). Es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C).