Predicción de complicaciones en biopsias pulmonares con IA

María Cribillés Pérez

Tutor: Francisco Herrera Triguero

22 de julio de 2025

UNIVERSIDAD DE GRANADA

Índice

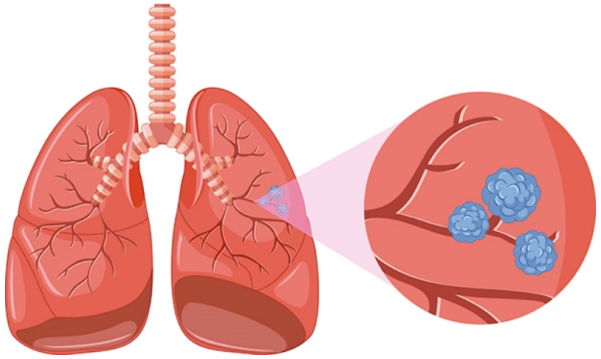
1. Introducción
   - Objetivos
2. Matemáticas
   - Señales
   - Radiómica
   - Optimización
   - Aprendizaje automático
   - Deep Learning
   - DML
3. Informática
   - Datos
   - Preprocesamiento
   - Modelos
   - Experimentación
   - Explicabilidad
4. Conclusiones

Contenido de Introducción

1. Introducción
   - Objetivos
2. Matemáticas
3. Informática
4. Conclusiones

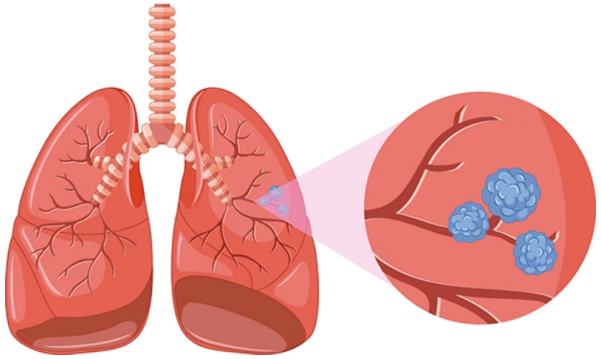
Introducción

- En 2020: **10 millones de fallecimientos** por cáncer.
- **2.2 millones** de nuevos diagnósticos de cáncer de pulmón.



Introducción

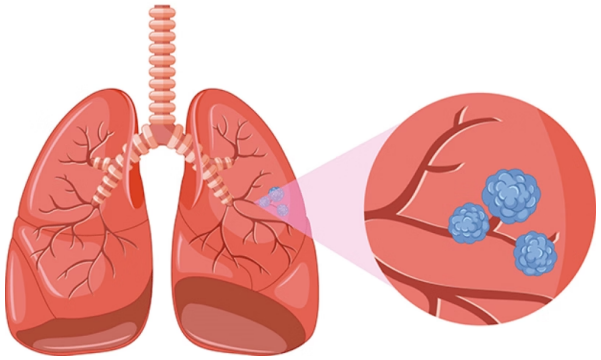
- En 2020: **10 millones de fallecimientos** por cáncer.
- **2.2 millones** de nuevos diagnósticos de cáncer de pulmón.



¿Cómo se diagnostica el cáncer de pulmón?

Introducción

- En 2020: **10 millones de fallecimientos** por cáncer.
- **2.2 millones** de nuevos diagnósticos de cáncer de pulmón.

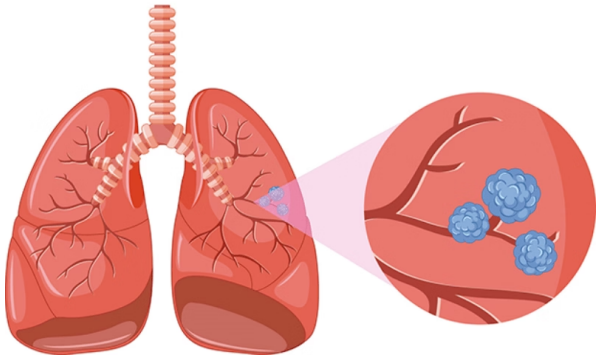


¿Cómo se diagnostica el cáncer de pulmón?

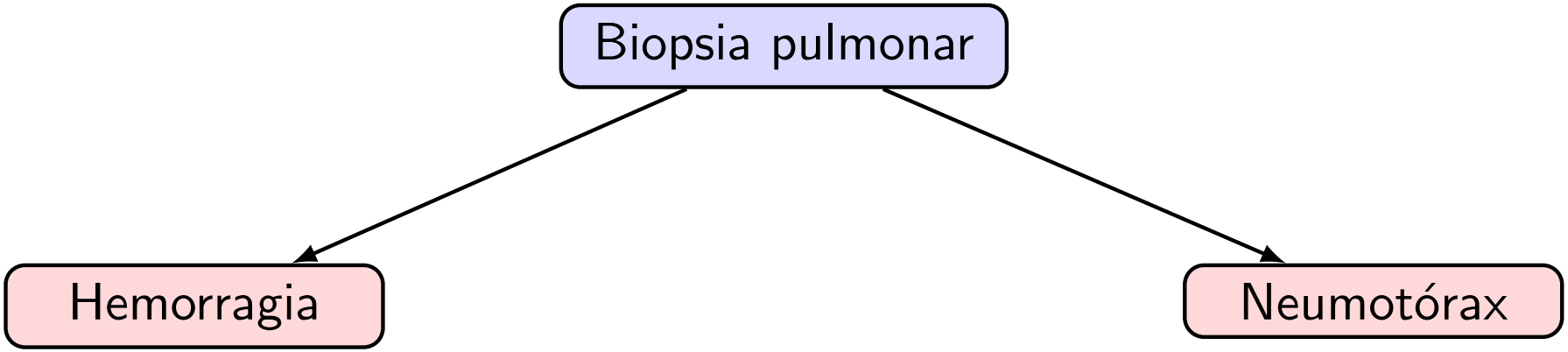
Biopsia pulmonar

Introducción

- En 2020: **10 millones de fallecimientos** por cáncer.
- **2.2 millones** de nuevos diagnósticos de cáncer de pulmón.

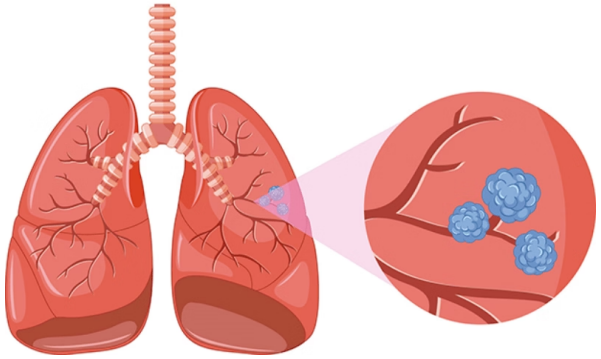


¿Cómo se diagnostica el cáncer de pulmón?



Introducción

- En 2020: **10 millones de fallecimientos** por cáncer.
- **2.2 millones** de nuevos diagnósticos de cáncer de pulmón.



¿Cómo se diagnostica el cáncer de pulmón?

Ausencia de literatura para predecir riesgos en biopsias.

Objetivos fundamentales:

- Estudiar los **fundamentos matemáticos** necesarios para la comprensión del problema abordado.
- **Analizar los datos** recibidos e interpretarlos adecuadamente para su posterior procesamiento.
- Elaborar **modelos** de predicción de complicaciones a partir de los datos de imagen y clínicos preprocesados.
- Desarrollar un **estudio experimental** que valide la efectividad de los modelos propuestos.
- **Analizar los resultados** obtenidos y aplicar técnicas de **explicabilidad** para extraer conclusiones relevantes para el equipo médico.

Contenido de Matemáticas

1. Introducción
2. Matemáticas
   - Señales
   - Radiómica
   - Optimización
   - Aprendizaje automático
   - Deep Learning
   - DML
3. Informática
4. Conclusiones

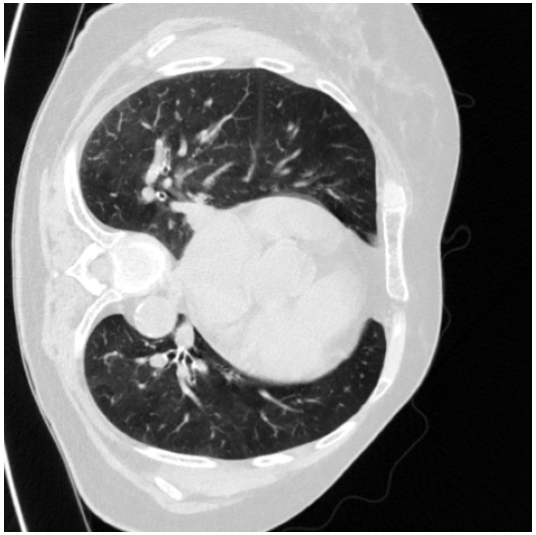
Teoría de preprocesamiento de señales

Definición de señal

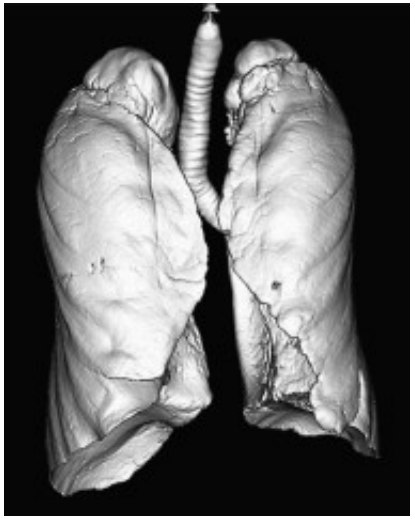
Una señal es una función que asigna valores reales y transmite información, pudiendo ser continua, $s : \mathbb{R}^n \to \mathbb{R}$, o discreta $s : \mathbb{Z}^n \to \mathbb{R}$.

Tipos de señales en imágenes médicas:

Imágenes 2D: señal discreta bidimensional



Imágenes 3D: señal discreta tridimensional



Transformada de Fourier continua

Definición (1D)

Sea $f \in L^1(\mathbb{R})$ y j la unidad imaginaria. Su transformada de Fourier es:

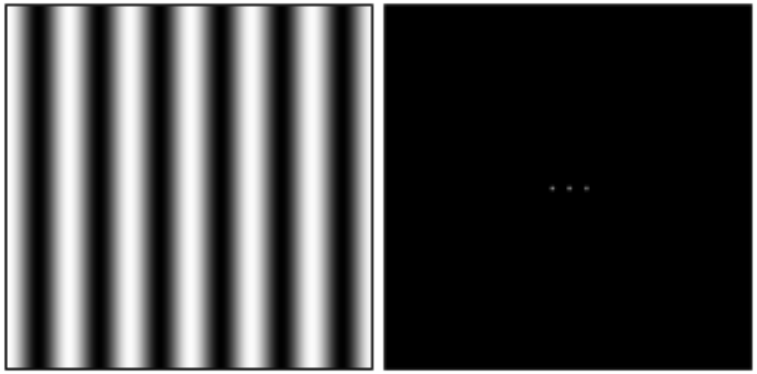
$$\mathcal{F}\{f\}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi j x \xi}\, dx \quad \xi \in \mathbb{R}.$$

Transformada inversa:

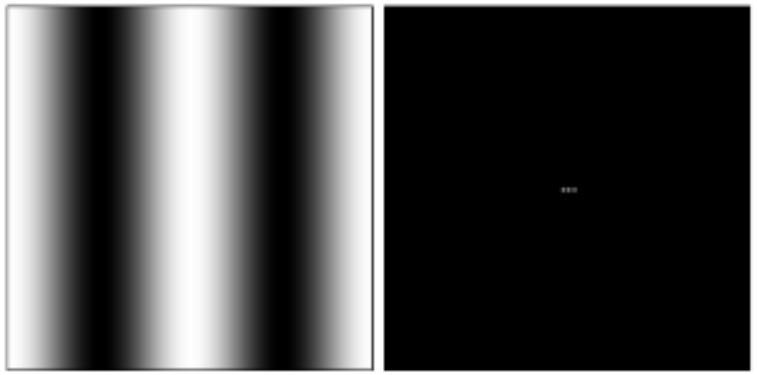
$$f(x) = \int_{-\infty}^{\infty} \mathcal{F}\{f\}(\xi), e^{2\pi j x \xi}\, d\xi.$$

Transformada de Fourier continua

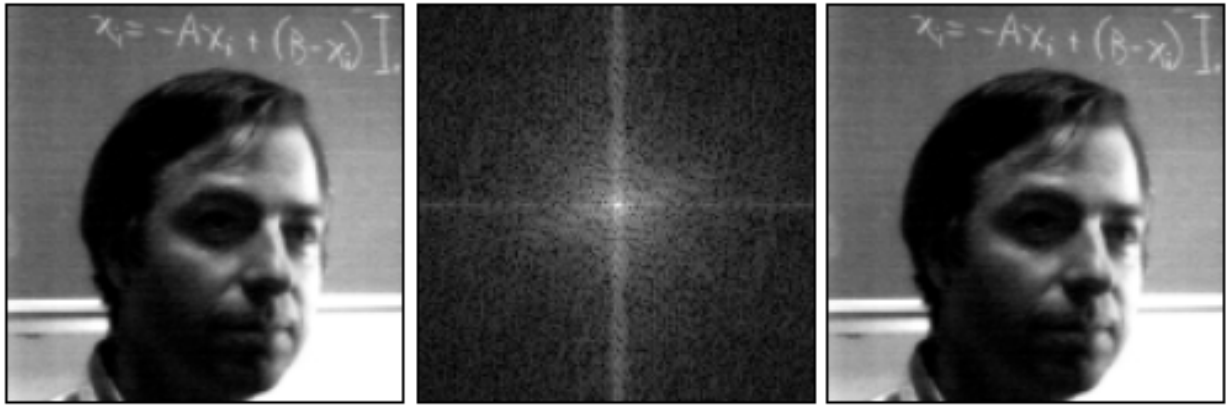
Señales



(a) Patrón sinusoidal vertical y su transformada de Fourier.



(b) Patrón sinusoidal con frecuencia espacial más baja y su transformada.



Transformada de Fourier directa e inversa de la imagen de entrada.

Operación de convolución

Definición general (continua)

Sean $f, g : \mathbb{R} \to \mathbb{R}$ dos funciones reales. La convolución se define como:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(a)\, g(t-a)\, da$$

donde f es la **entrada**, g es el **núcleo** o *kernel/filtro* y el resultado $(f * g)$ es el **mapa de características** o *feature map*.

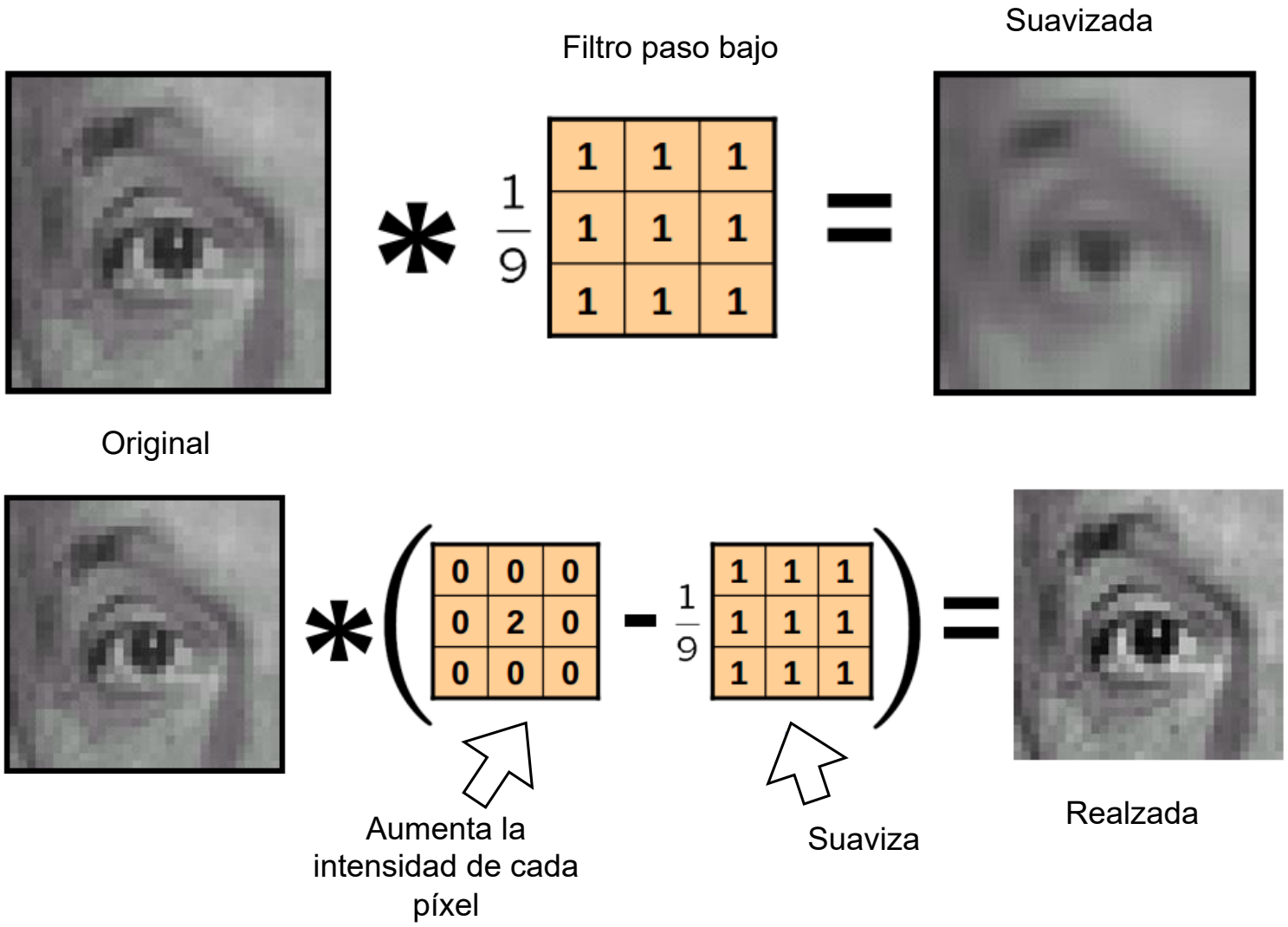
Convolución discreta (1D)

Para señales discretas:

$$(f * g)(t) = \sum_{a=-\infty}^{\infty} f(a)\, g(t-a).$$

Operación de convolución

Señales



$$* \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} =$$

$$* \left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \right) =$$

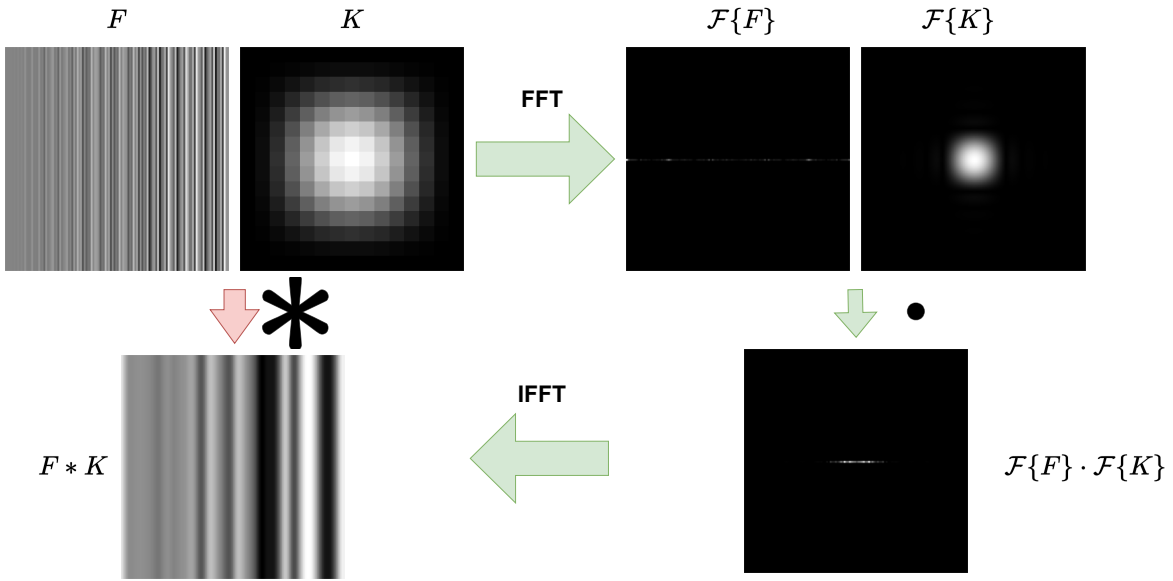
Teorema fundamental de la convolución

Teorema de la convolución

La transformada de Fourier convierte convoluciones en el dominio espacial en multiplicaciones en el dominio de la frecuencia, y viceversa:

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\}$$

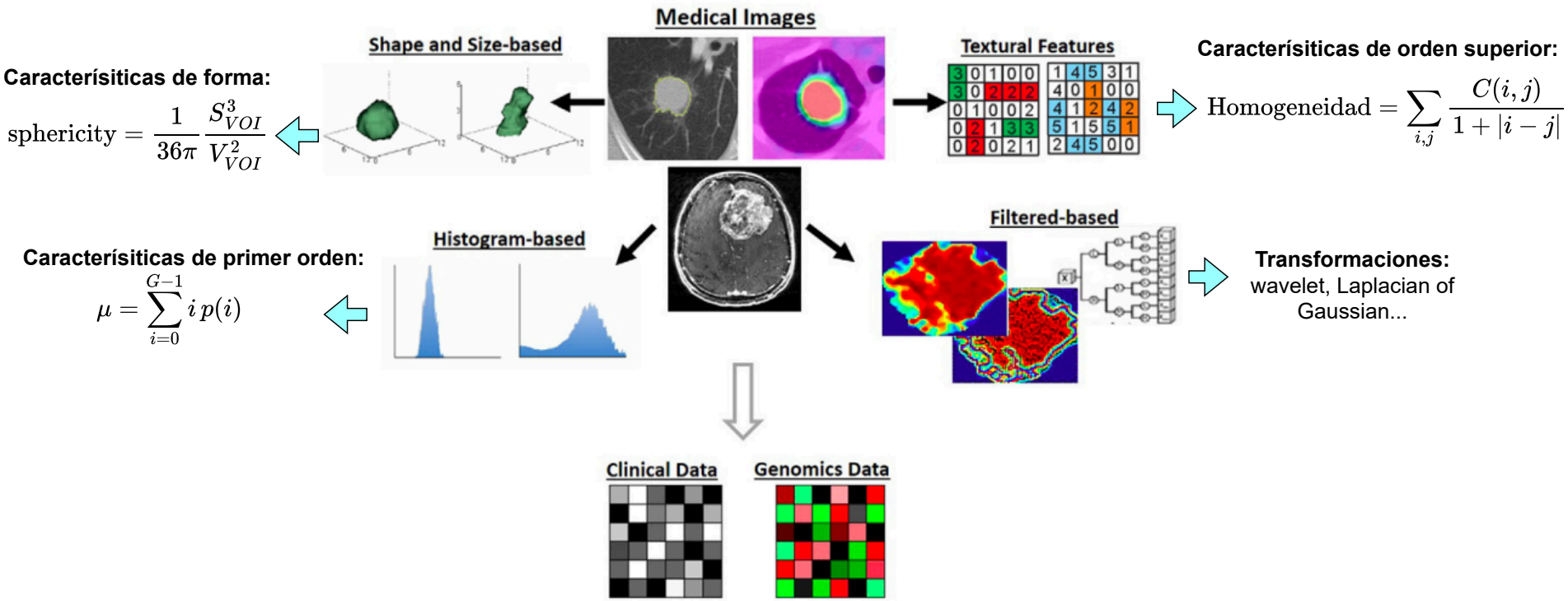
$$\mathcal{F}\{f \cdot g\} = \mathcal{F}\{f\} * \mathcal{F}\{g\}$$



Radiómica

Definición

Campo de la medicina de precisión que extrae automáticamente grandes cantidades de características cuantitativas de imágenes o volúmenes médicos.



Optimización y convexidad

Definición: Descenso de gradiente

Método iterativo para minimizar una función objetivo diferenciable avanzando en la dirección opuesta al gradiente:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

donde η es la tasa de aprendizaje.

Descenso de gradiente con proyecciones

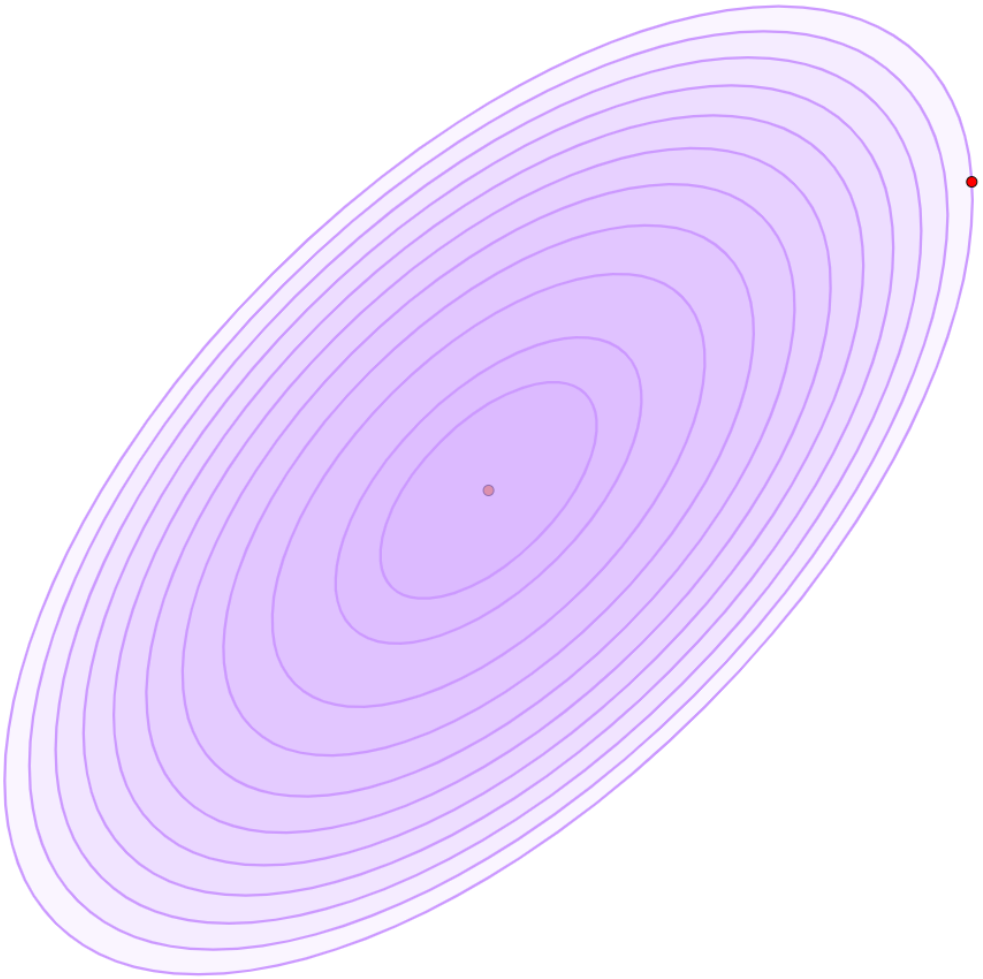
Cuando el dominio de la función está restringido a un conjunto convexo C, el descenso de gradiente estándar no garantiza que los puntos generados sigan siendo viables. Se proyecta el punto sobre C en cada iteración:

$$x_{t+1} = P_C(x_t - \eta \nabla f(x_t))$$

donde P_C es la proyección sobre el conjunto C.

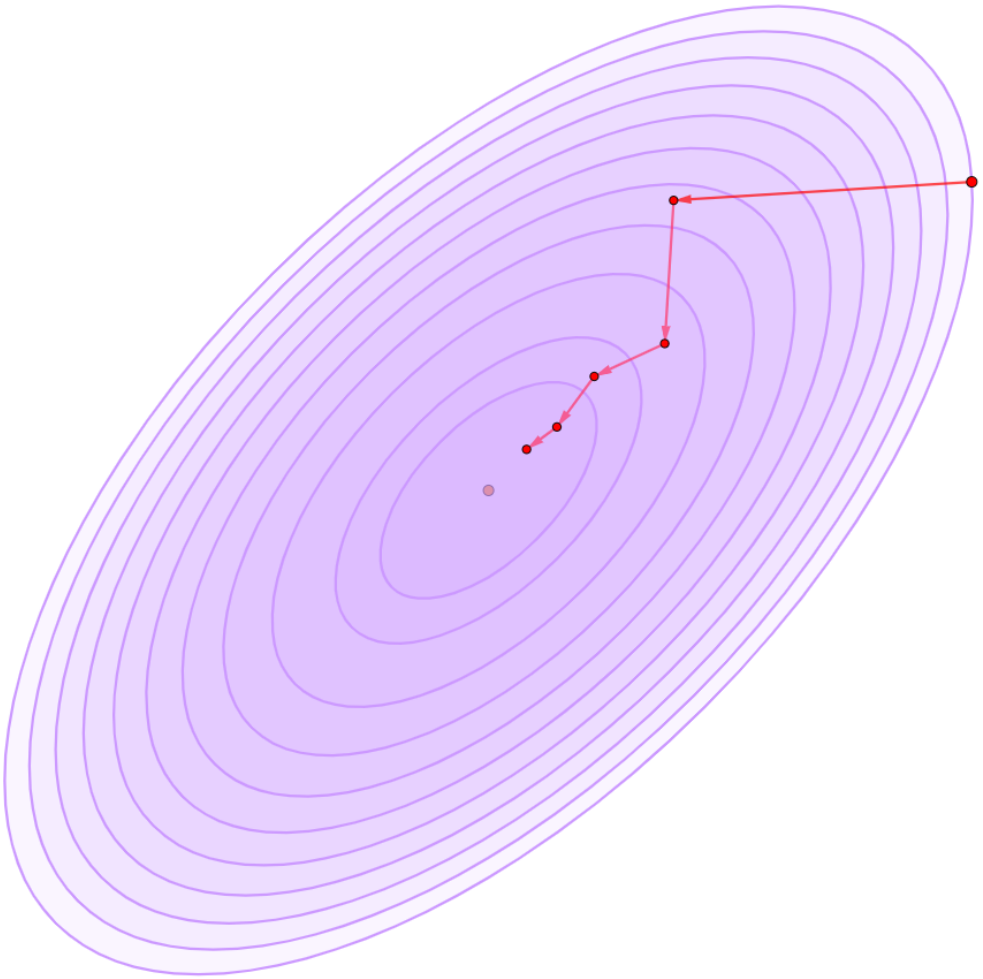
Optimización y convexidad

Optimización



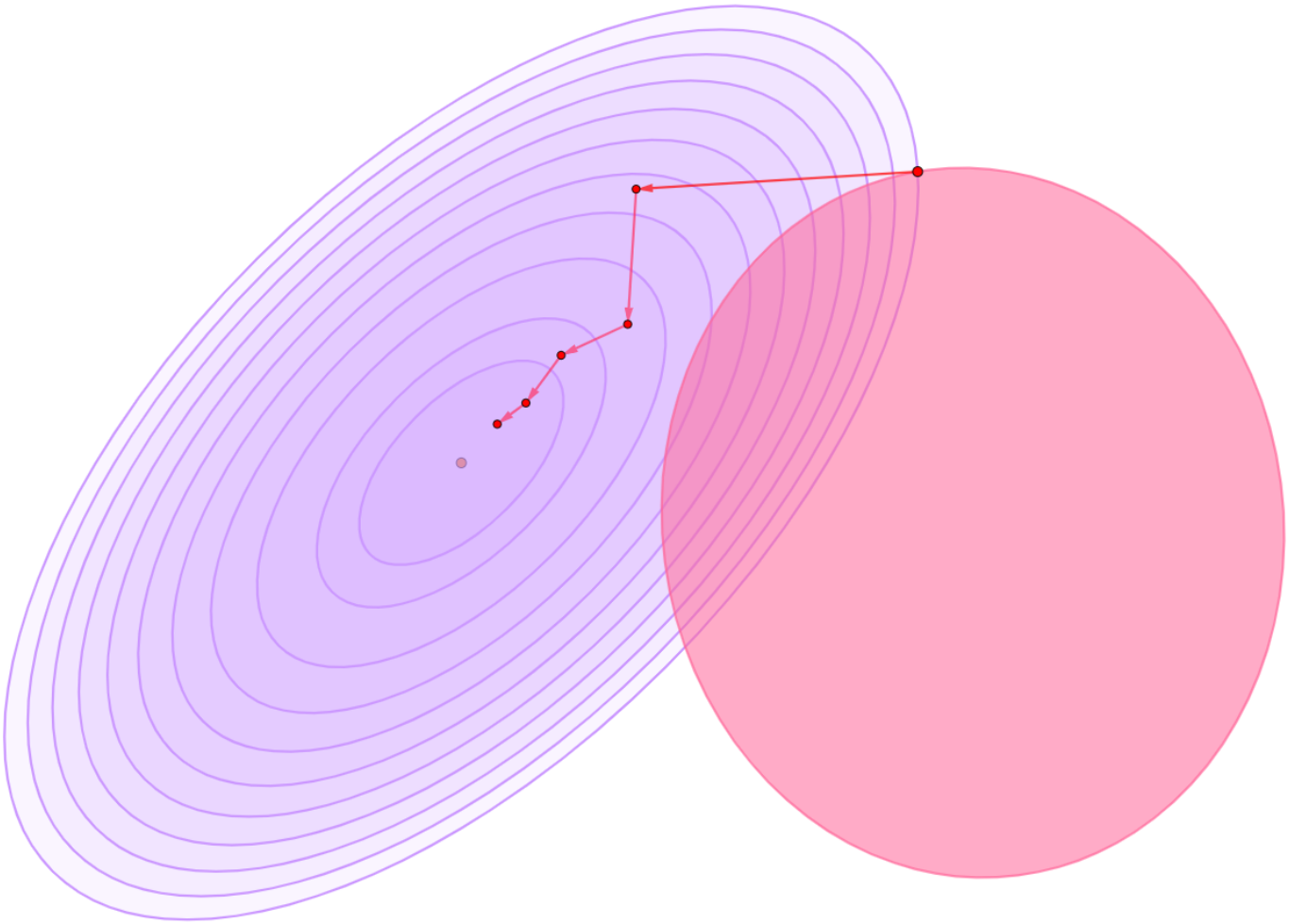
Optimización y convexidad

Optimización



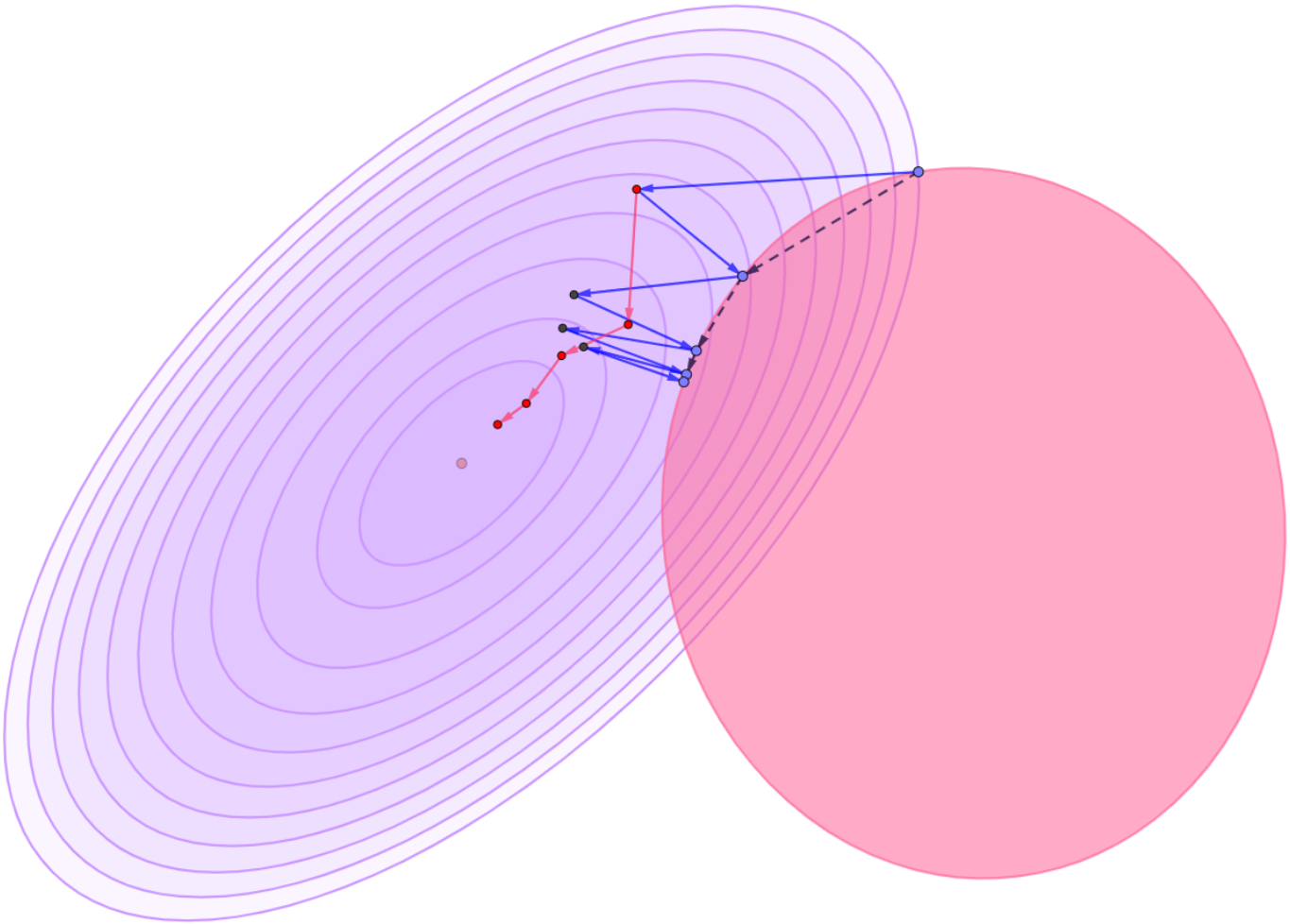
Optimización y convexidad

Optimización



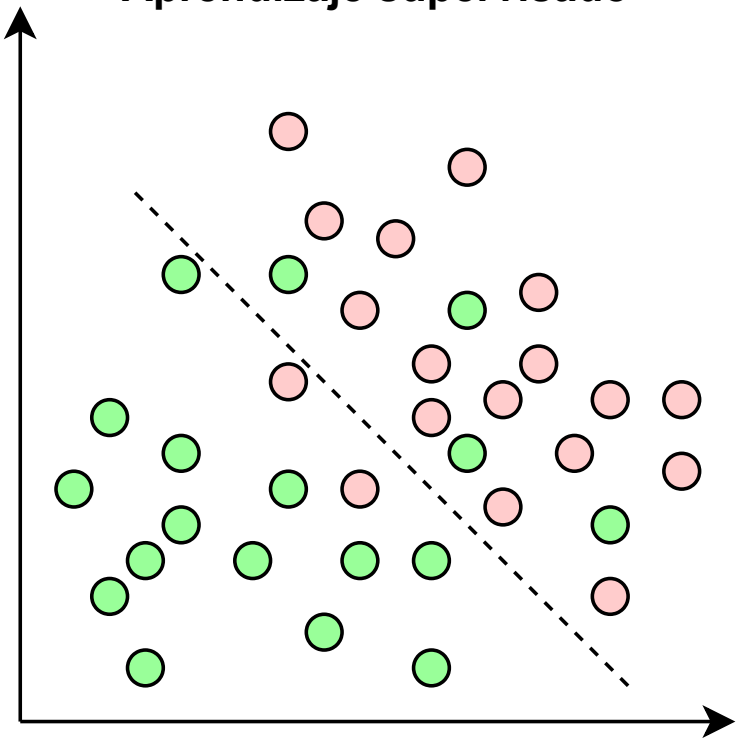
Optimización y convexidad

Optimización

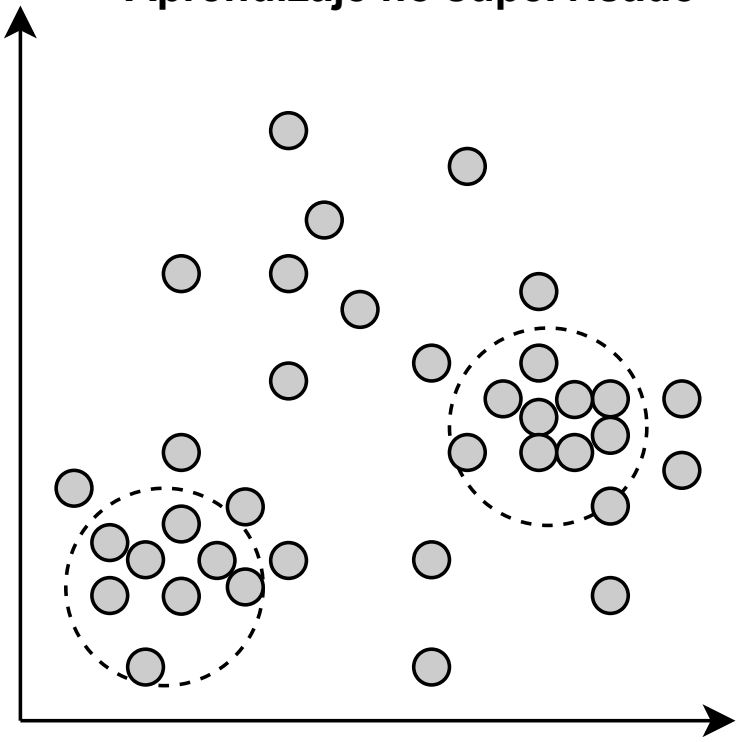


Tipos de aprendizaje automático

Aprendizaje supervisado



Aprendizaje no supervisado



Regla de predicción: $h : \mathcal{X} \to \mathcal{Y}$

Aprende distribución de probabilidad: $p(x)$

Conjunto de entrenamiento:

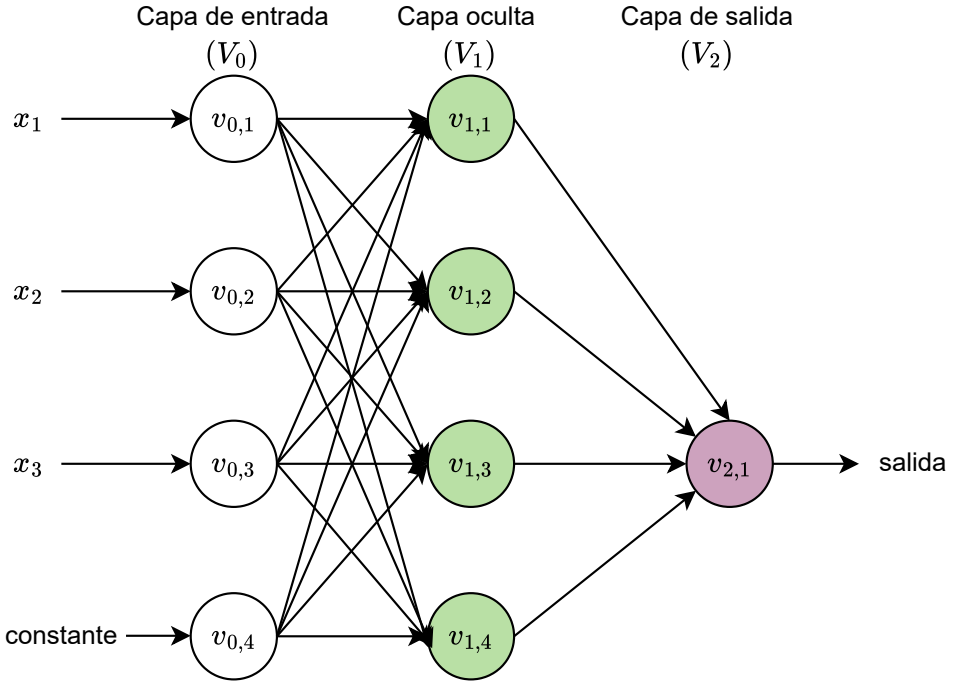
$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \mathcal{Y}$$

Red neuronal artificial (RNA)

Definición

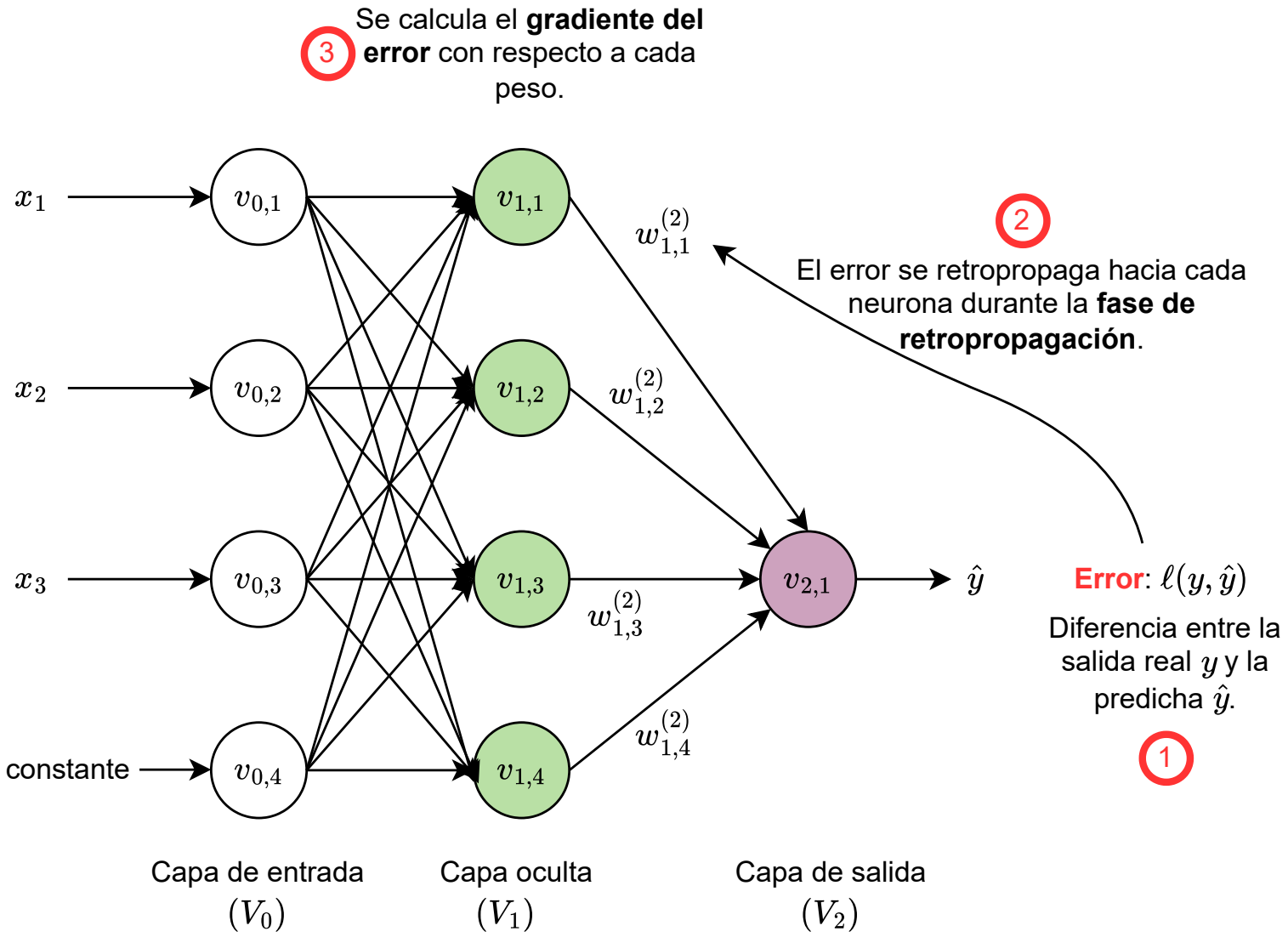
Una **RNA** define una función $y = f(x; \theta)$ como composición de transformaciones:

$$f(x) = f^{(L)}(f^{(L-1)}(\dots f^{(1)}(x)))$$



Algoritmo de backpropagation

Deep Learning



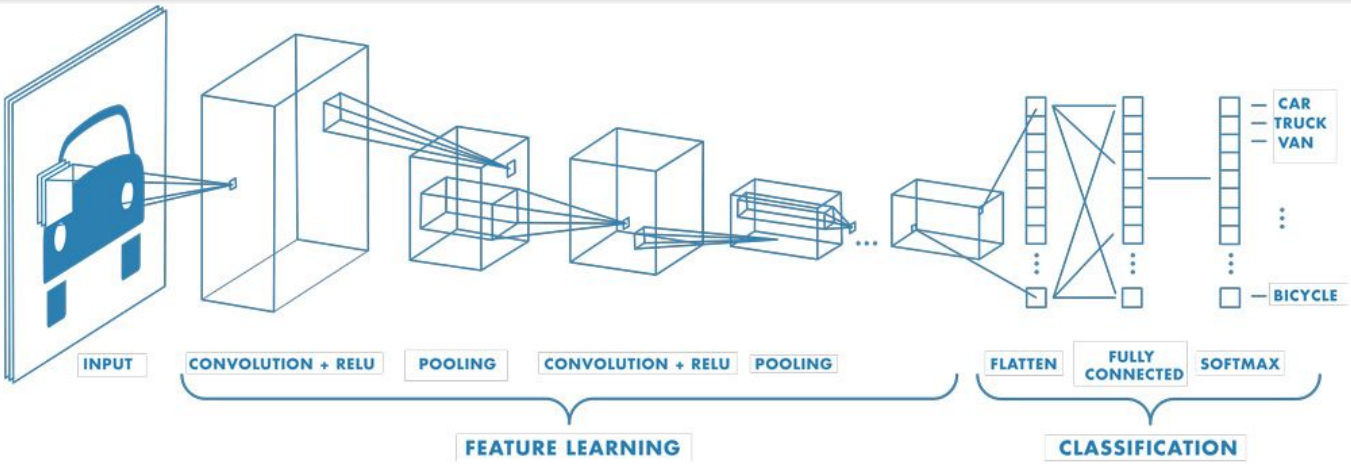
Transfer learning y redes convolucionales (CNNs)

Transfer learning

El **aprendizaje por transferencia** consiste en reutilizar un modelo previamente entrenado en una tarea (fuente) para resolver una nueva tarea (destino), ajustando sus últimas capas o realizando *fine-tuning*.

Redes neuronales convolucionales (CNNs)

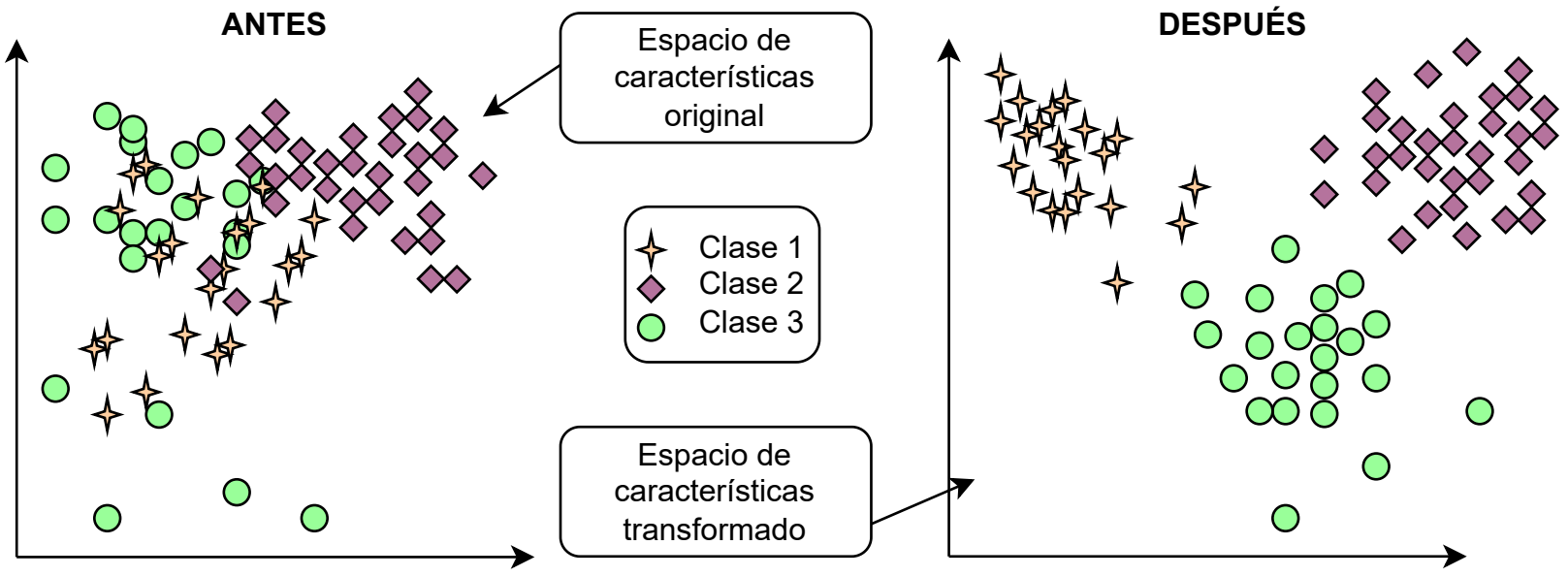
Tipo de RNA que utiliza capas convolucionales para identificar patrones espaciales.



Aprendizaje de métricas de distancia

Definición general

El **Aprendizaje de Métricas de Distancia (DML)** busca aprender una transformación del espacio de características que mejore la medida de similitud entre instancias, de forma que elementos similares queden más cerca y elementos distintos más alejados según la nueva métrica.



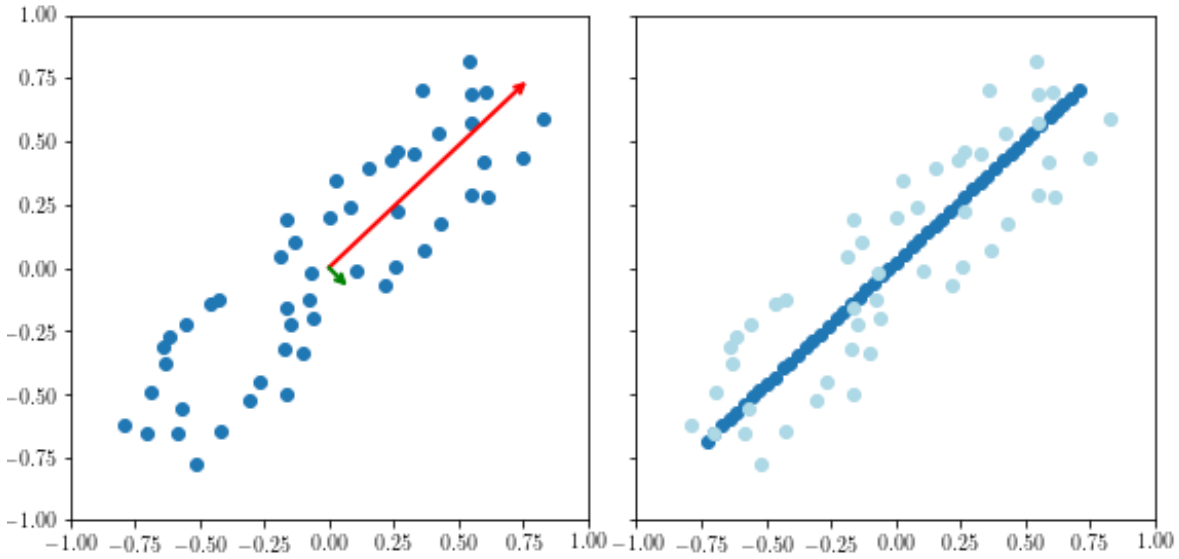
Aprendizaje de métricas de distancia

- **PCA (Análisis de Componentes Principales):** reducción de dimensionalidad no supervisada.
- **NCA (Neighbourhood Components Analysis):** mejora el rendimiento de un clasificador k-NN supervisadamente.
- **LMNN (Large Margin Nearest Neighbors):** utiliza función de pérdida más geométrica que NCA: acerca a ejemplos de una misma clase y aleja de diferentes clases.

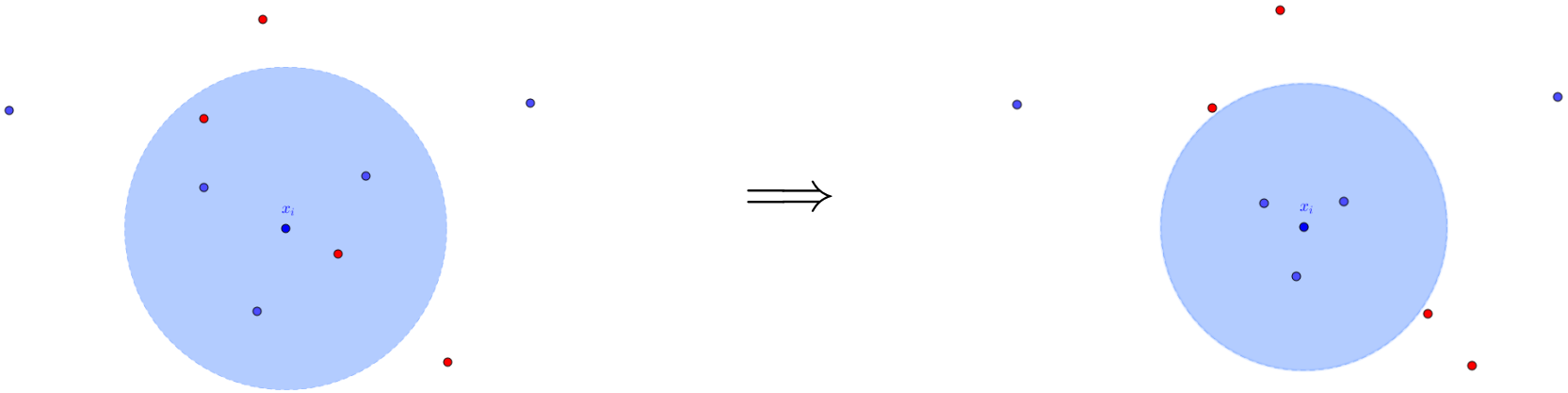
Aprendizaje de métricas de distancia

DML

PCA



LMNN



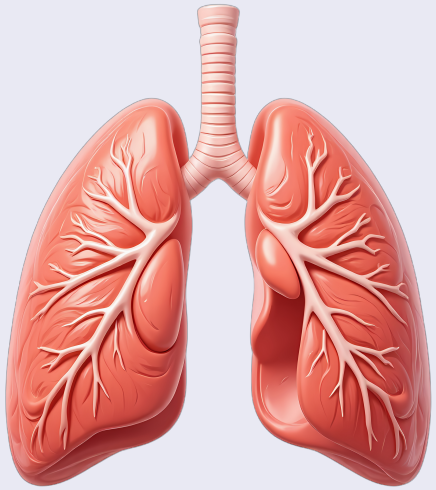
Contenido de Informática

1. Introducción
2. Matemáticas
3. Informática
   - Datos
   - Preprocesamiento
   - Modelos
   - Experimentación
   - Explicabilidad
4. Conclusiones

Datos

Recogida en lotes $\Longrightarrow$ **Total**: 125 pacientes del Hospital Universitario Clínico San Cecilio de Granada

Volúmenes 3D



Datos clínicos

Datos

Recogida en lotes $\Longrightarrow$ **Total**: 125 pacientes del Hospital Universitario Clínico San Cecilio de Granada

Volúmenes 3D

- Imágenes TC torácicas.
- Formato: DICOM.
- Tamaño: (768,768,x) y (512,512,x).
- Diferente número de slices.

Datos clínicos

- Edad y sexo.
- Factores de riesgo: tabaquismo, obesidad, etc.
- Patología pulmonar: neumonía, enfisema, etc.
- Etiqueta: complicación sí/no.
- Otras variables post-biopsia.
- Anonimizados.

Preprocesamiento de los datos

Preprocesamiento de volúmenes médicos

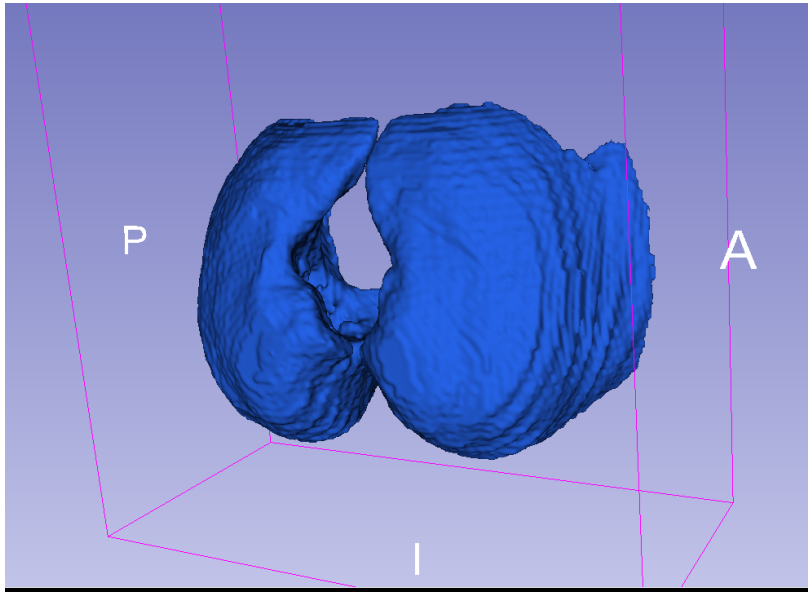
- Ventana pulmonar en unidades **Hounsfield**.
- Redimensionamiento a tamaño fijo de **slices** y tamaño de imagen.
- **Segmentación** del pulmón.
- **Data augmentation**: leve rotación, escalado y traslaciones aleatorias.

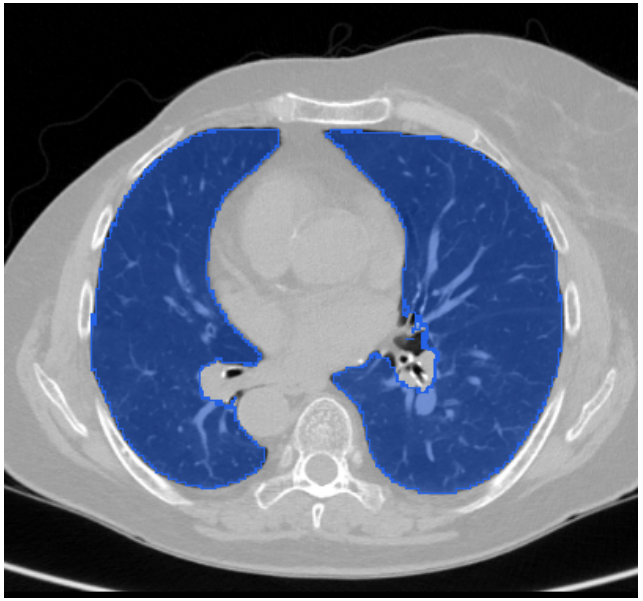
Preprocesamiento de variables clínicas

- **Codificación binaria** de variables categóricas (ej.: sexo, complicación, factores de riesgo).
- Eliminación de **fallos ortográficos** y **sinónimos** médicos.
- Escalado **MinMax** de la variable edad.

Preprocesamiento de los datos

Preprocesamiento





Modelado 2D

Modelo empleado

EfficientNetV2

- Requieren menos **capacidad de cómputo**.
- Cortes individuales $\Longrightarrow$ **pérdida de información**.
- Más **desbalanceo** entre clases.
- Mayor variabilidad e **inestabilidad**: corte aislado $\neq$ paciente completo.

Modelado 2D

Modelo empleado

Efficier V2

- Requieren menos **ca ad de uto**.
- Cortes individua → **pérdida d mación**.
- Más **desbala** entre clases.
- Mayor va dad e **inestabilidad**: co ado ≠ paciente comple

Modelado 3D y estrategias de entrenamiento

Modelos empleados

ResNet3D y **DenseNet3D**

- **Función de pérdida:** CrossEntropyLoss.
- **Optimizador:** Adam + weight decay.
- **Regularización:** Early Stopping, scheduler de *learning rate* / LR fijo y Dropout.
- **Batch size:** variado 2, 4, 6, 8, 16 y 32.

Experimentos realizados:

- Modelo base con volúmenes 3D.
- Modelo multimodal (volúmenes + datos clínicos).
- Preentrenamiento MedMNIST3D + *Fine-tuning*: congelado y descongelado.

Experimentación con radiómica

Extracción de características (PyRadiomics)

- Original: 107 características, sin transformar la imagen.
- Extended: 1967 características con las transformaciones.

Modelos empleados

LightGBM, Gradient Boosting, Random Forest, **KNN**, árboles de decisión, SVM y regresión logística.

Preprocesamiento	**Original**	**Extendido**
Dimensión inicial	107	1967
Sin redundantes	89	1382
Variance Threshold	79	588
PCA 99 %	23	68
PCA 95 %	14	33
PCA 90 %	10	20

Reducción de dimensionalidad.

Aprendizaje de métricas de distancia (DML)

Objetivo general

Aprender una distancia a partir de los datos.

Experimentos con radiómica

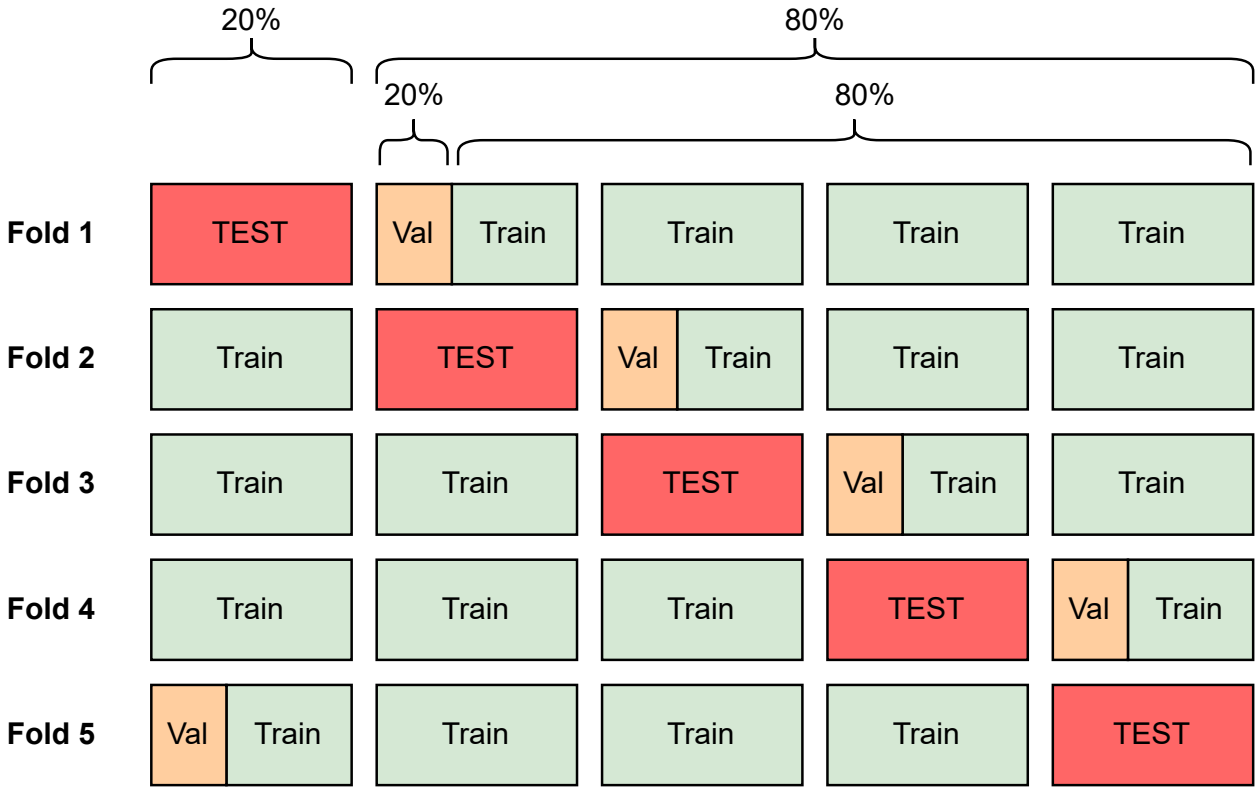
- Sin datos clínicos ni DML.
- Con datos clínicos y sin DML.
- Sin datos clínicos y con DML.
- Con datos clínicos y con DML.

Validación

Experimentación

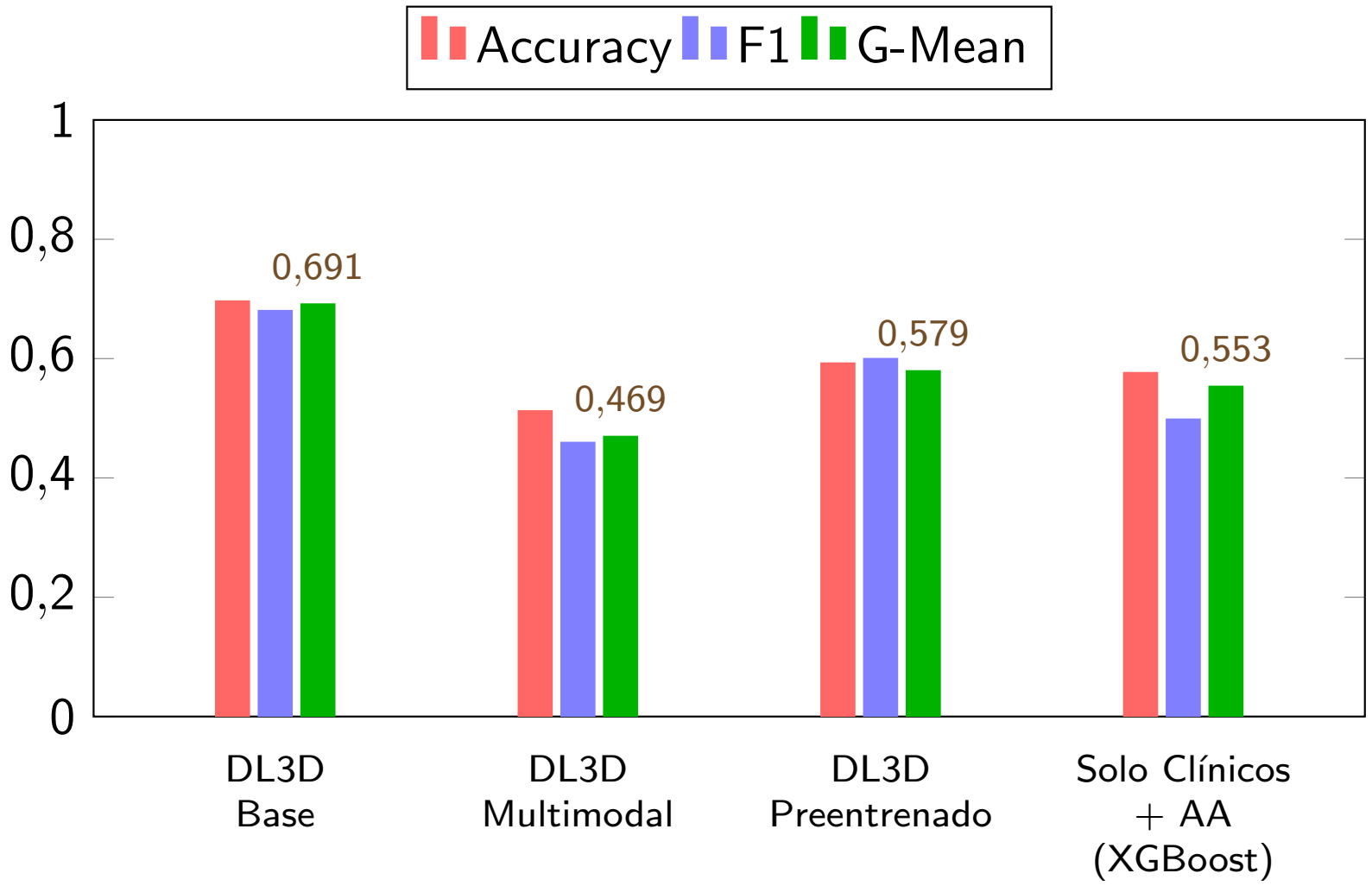
Parámetros comunes de evaluación

- **Validación cruzada estratificada** (k=5) y **Early stopping**.
- Métricas: **Accuracy**, F1-score, TPR, TNR y **G-Mean**.



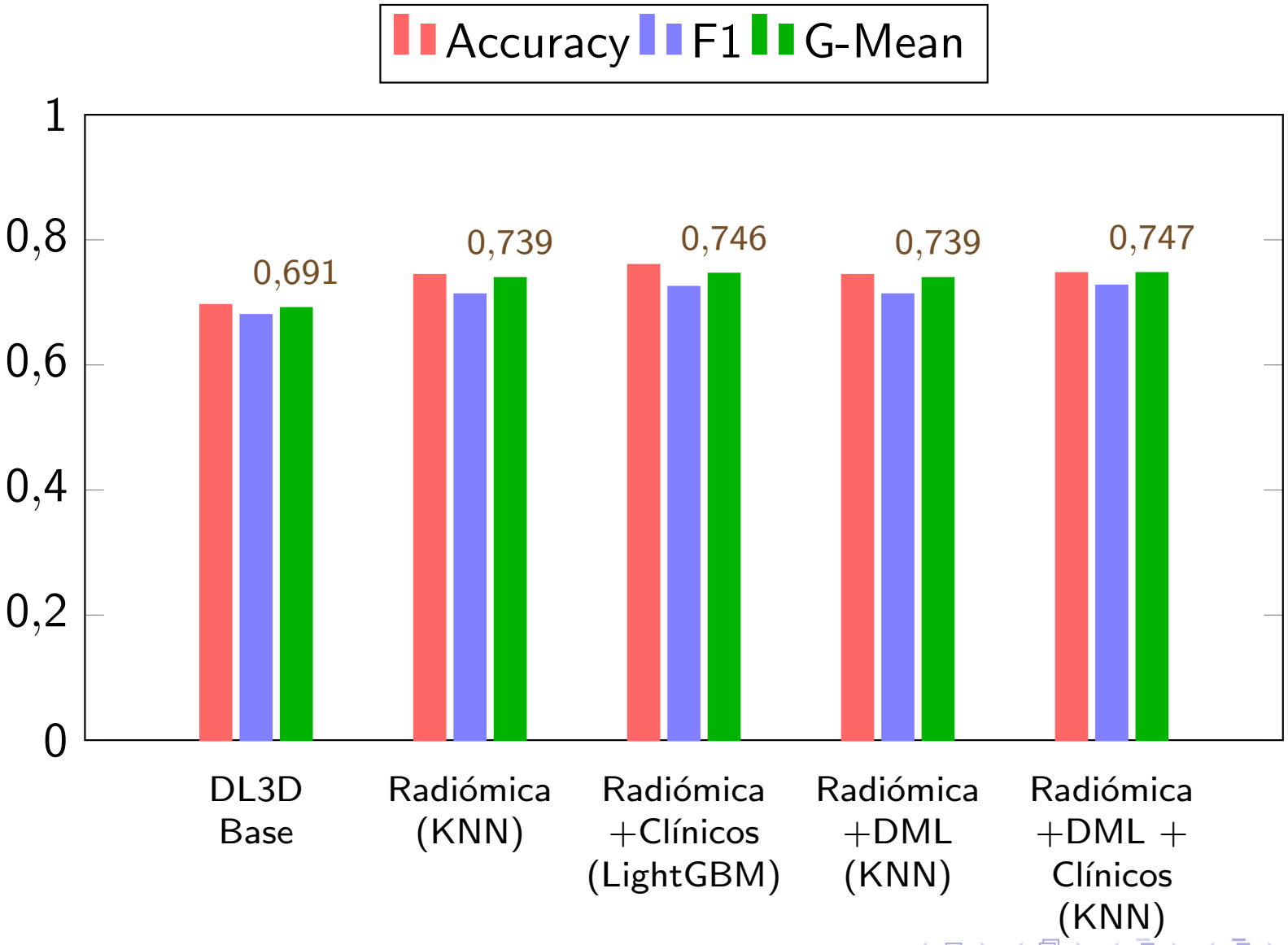
Comparación de métricas por modelo (mejor resultado)

Experimentación



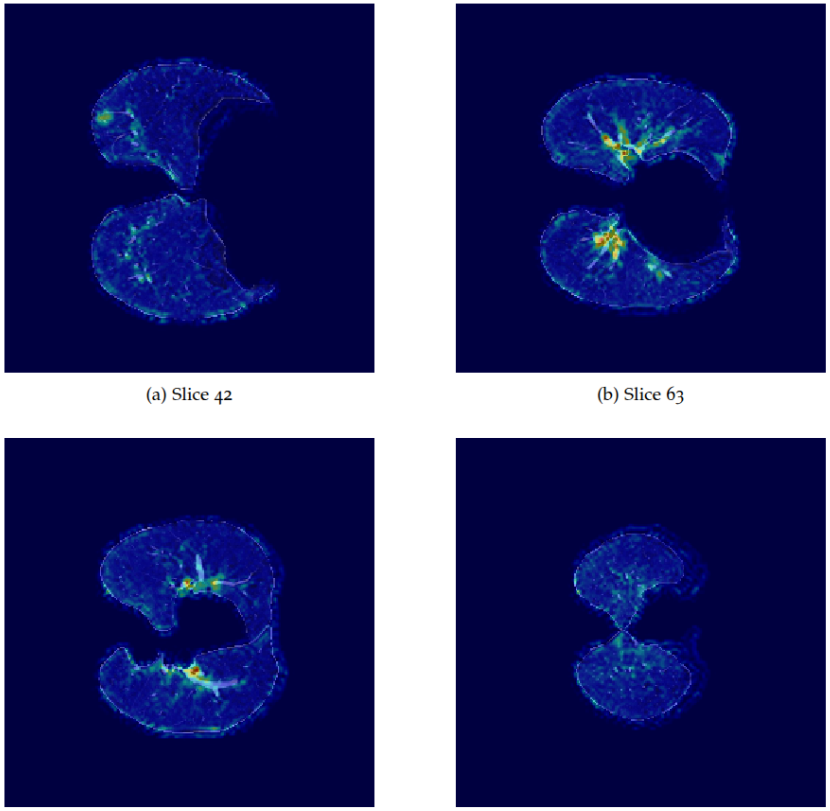
Comparación de métricas por modelo (mejor resultado)

Experimentación



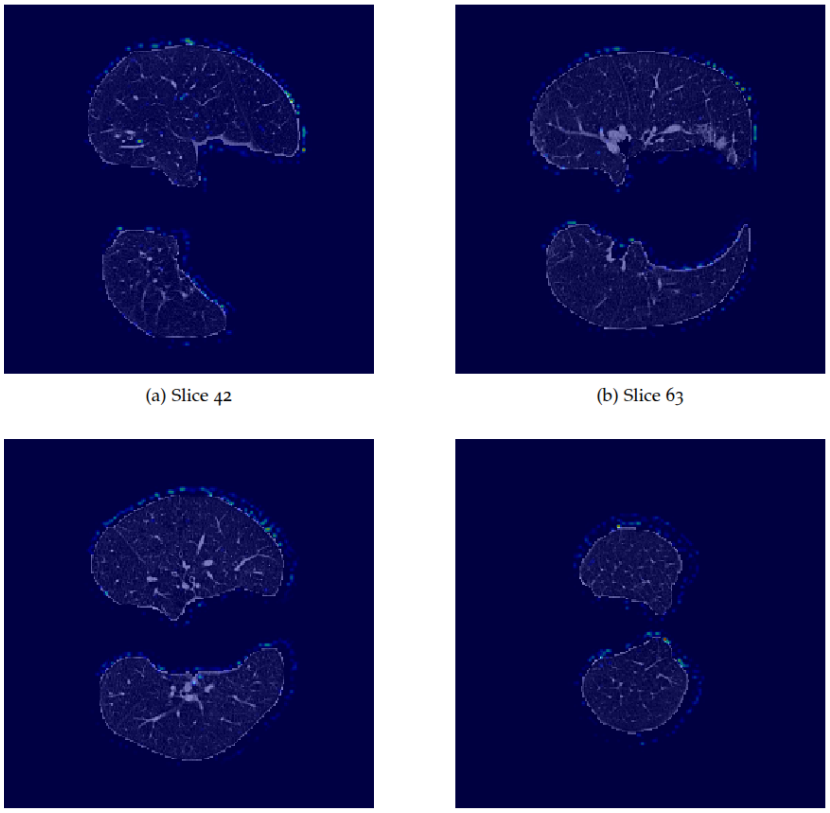
Grad-CAM

Explicabilidad



(a) Slice 42 (b) Slice 63 (c) Slice 84 (d) Slice 105

Instancia con complicación predicha correctamente

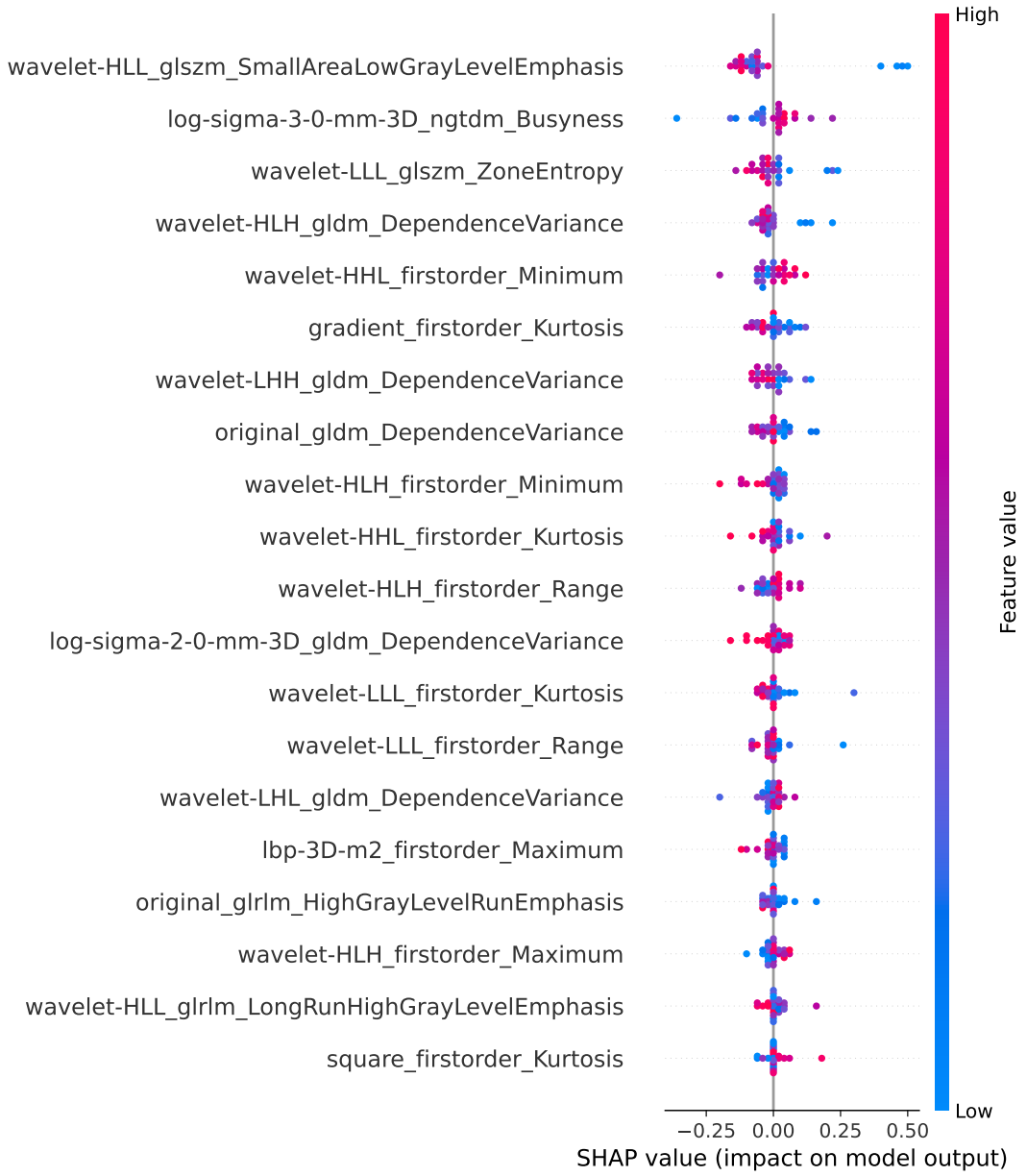


(a) Slice 42 (b) Slice 63 (c) Slice 84 (d) Slice 105

Instancia con complicación predicha incorrectamente

SHAP

Explicabilidad



Contenido de Conclusiones

1. Introducción
2. Matemáticas
3. Informática
4. Conclusiones

Conclusiones y líneas futuras

Resumen de resultados

Multimodal $<$ **Clínicos** $<$ **Fine-tuning** $<$ **DL 3D** $<$ **Radiómica + DML**

Dificultades

Trabajar con volúmenes 3D, disponibilidad de los datos, desbalanceo de clases...

Líneas de trabajo futuro

1. Predicción del **tipo de complicación** (multietiqueta).
2. Segmentación hacia **estructuras anatómicas clave** (tumores, vasos).
3. **Aprendizaje federado** entre hospitales para aumentar datos sin comprometer privacidad.
4. Estudiar el **impacto de variables técnicas** como el modelo de TC.
5. Incorporar **mecanismos de atención** y mejorar explicabilidad.
6. Investigar **estrategias semisupervisadas** con datos no etiquetados.

Bibliografía fundamental

- Goodfellow, I., Bengio, Y., Courville, A. y Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT Press. Utilizada para el estudio de la teoría de optimización y los fundamentos de Deep Learning.
- Shalev-Shwartz, S., y Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. Utilizada para los fundamentos de aprendizaje automático y fundamentos del aprendizaje de métricas de distancias.
- Un artículo fundamental para el estudio de la radiómica y su base matemática: de Vaucleroy, N., Macq, B., De Vleeschouwer, C., y Léger, J. *Mathematical morphology applied to Radiomics*.
- Artículos relacionados con la parte práctica aplicada a medicina.

Muchas gracias por su atención

María Cribillés Pérez