Selección de Muestra para la Detección Automática de Nódulos Pulmonares en CT

Andrés Montoya

14 de noviembre de 2025

1. Diseño Metodológico y Selección de Muestra

1.1. Definición de la población y marco muestral

La población objeto de estudio está compuesta por estudios de tomografía computarizada (CT) de tórax provenientes de distintas instituciones médicas, adquiridos bajo protocolos de diagnóstico pulmonar. El marco muestral estará conformado por la totalidad de estudios disponibles en formato `NIfTI`, anonimizados y con metadatos que describan las condiciones de adquisición (fabricante, grosor de corte, edad y sexo del paciente, entre otros).

El objetivo principal es desarrollar y validar una red neuronal convolucional (CNN) capaz de detectar la **presencia o ausencia de nódulos pulmonares** en imágenes CT.

1.2. Estimación de parámetros poblacionales

Se define la prevalencia de nódulos pulmonares en la población de referencia como p, la cual puede ser estimada a partir de literatura o estudios piloto. Asimismo, se establece un nivel de confianza del 95 % y un margen de error admisible E. El valor crítico correspondiente al nivel de confianza se denota como Z, con $Z = 1{,}96$ para una distribución normal estándar.

1.3. Cálculo del tamaño de muestra para estimar una proporción

El tamaño de muestra requerido para estimar una proporción poblacional (por ejemplo, la prevalencia de nódulos) se calcula mediante la siguiente expresión:

$$n = \frac{Z^2\, p(1-p)}{E^2} \tag{1}$$

donde:

- n es el tamaño de muestra requerido,

- Z es el valor crítico para el nivel de confianza seleccionado,
- p es la proporción esperada (prevalencia),
- E es el margen de error máximo tolerado.

Por ejemplo, si se asume una prevalencia esperada de $p = 0{,}05$ (5 %) y un error máximo $E = 0{,}05$, entonces:

$$n = \frac{1{,}96^2 \times 0{,}05 \times (1 - 0{,}05)}{0{,}05^2} = 73$$

Por tanto, se requieren aproximadamente 73 estudios para estimar la prevalencia con una confianza del 95 % y un error de ±5 %.

1.4. Cálculo del tamaño de muestra para sensibilidad y especificidad

Dado que el objetivo es evaluar el rendimiento diagnóstico del modelo, se deben estimar también las muestras necesarias para calcular la *sensibilidad* (Se) y la *especificidad* (Sp) del clasificador.

El número mínimo de casos positivos necesarios para estimar la sensibilidad con un margen de error d se obtiene mediante:

$$n_{pos} = \frac{Z^2 \, \mathrm{Se}(1 - \mathrm{Se})}{d^2} \tag{2}$$

De forma análoga, para la especificidad:

$$n_{neg} = \frac{Z^2 \, \mathrm{Sp}(1 - \mathrm{Sp})}{d^2} \tag{3}$$

Ejemplo: si se espera una sensibilidad de 0,9 con una precisión de ±0,05 al 95 % de confianza, se requiere:

$$n_{pos} = \frac{1{,}96^2 \times 0{,}9 \times (1 - 0{,}9)}{0{,}05^2} = 139$$

Por tanto, se necesitan aproximadamente 139 estudios positivos (con nódulos) para una estimación confiable de la sensibilidad.

1.5. Estrategia de muestreo y balance de clases

Dada la baja prevalencia de nódulos pulmonares en estudios CT generales, se implementará un **muestreo estratificado caso–control**, donde se seleccionarán de manera balanceada:

- Casos positivos (estudios con presencia de nódulos),
- Casos negativos (estudios sin nódulos),

- Diversas fuentes institucionales y protocolos de adquisición.

Para el conjunto de entrenamiento del modelo, se permitirá el sobre-muestreo de la clase minoritaria (casos positivos) y la aplicación de técnicas de *data augmentation* (rotaciones, desplazamientos, variaciones de intensidad). Los conjuntos de validación y prueba mantendrán una distribución representativa de la prevalencia real para una evaluación fiel del desempeño del modelo.

1.6. Estratificación y representatividad

La estratificación y selección de la muestra se realizará con base en la información contenida en el archivo `metadata.csv`, que consolida los metadatos técnicos y demográficos asociados a cada estudio. Dicho archivo incluye variables tales como:

- **Identificación y descripción del estudio:** `VolumeName`, `SeriesDescription`, `StudyDate`.
- **Características del equipo y adquisición:** `Manufacturer`, `ManufacturerModelName`, `ConvolutionKernel`, `FilterType`, `CTDIvol`.
- **Geometría y parámetros técnicos:** `Rows`, `Columns`, `ReconstructionDiameter`, `ZSpacing`, `XYSpacing`.
- **Datos del paciente:** `PatientSex`, `PatientAge`, `PatientPosition`.

La selección se llevará a cabo mediante un proceso de depuración, filtrado y estratificación que garantice representatividad en los siguientes ejes:

- Fabricante y modelo del escáner,
- Sexo y grupos etarios del paciente,
- Institución de procedencia (si está disponible),
- Protocolo de reconstrucción y grosor de corte,
- Calidad y completitud de los datos registrados.

De esta forma, se busca asegurar una muestra técnicamente heterogénea pero estadísticamente equilibrada, que refleje la diversidad real de condiciones presentes en los estudios CT de tórax.

1.7. Procedimiento general de selección y entrenamiento

El procedimiento metodológico completo se detalla en los siguientes pasos:

1. Definir la población y construir el marco muestral con metadatos.
2. Estimar la prevalencia esperada y calcular el tamaño de muestra necesario según las ecuaciones anteriores.
3. Aplicar muestreo estratificado garantizando representatividad.
4. Recolectar y anonimizar los estudios seleccionados.
5. Etiquetar los casos por radiólogos expertos siguiendo un protocolo estandarizado.
6. Dividir los datos en conjuntos de entrenamiento, validación y prueba (por ejemplo, 70/15/15), asegurando un número mínimo de casos positivos en cada subconjunto.
7. Implementar el preprocesamiento (resampling, normalización, segmentación pulmonar).
8. Entrenar el modelo de detección.
9. Evaluar el desempeño en el conjunto de prueba representativo y reportar sensibilidad, especificidad y área bajo la curva ROC.

Este enfoque garantiza una selección de datos sólida, una estimación estadísticamente válida de las métricas diagnósticas y un entrenamiento reproducible del modelo de detección de nódulos pulmonares.

1.8. Evaluación mediante la Curva ROC y el AUC

La **curva ROC** (Receiver Operating Characteristic) es una herramienta fundamental para evaluar el rendimiento diagnóstico de modelos binarios, como aquel destinado a detectar la presencia o ausencia de nódulos pulmonares.

Esta curva representa gráficamente la relación entre la *tasa de verdaderos positivos* (sensibilidad) y la *tasa de falsos positivos* a medida que se modifica el umbral de decisión del modelo.

Un modelo ideal se ubica en la esquina superior izquierda del gráfico (sensibilidad = 1, falsos positivos = 0), mientras que un clasificador aleatorio se sitúa sobre la diagonal principal.

El área bajo la curva, o **AUC (Area Under the Curve)**, cuantifica la capacidad del modelo para discriminar entre casos positivos y negativos:

$$0 \leq \text{AUC} \leq 1$$

- AUC = 1,0: rendimiento perfecto,
- AUC = 0,5: comportamiento aleatorio,
- AUC < 0,5: desempeño peor que el azar.

En el contexto de este estudio, la curva ROC permitirá determinar el umbral de probabilidad óptimo para la clasificación de estudios con y sin nódulos pulmonares, asegurando un balance adecuado entre sensibilidad y especificidad.