Miniproyecto Entrega 1

Natalia Ortiz	María Camila Santamaría	Eduardo Herrera Alba
202012954	202015359	201912865
n.ortizv	m.santamariaa	ej.herreraa

1. Revisión de estadísticas

1.1. ¿Porque es necesario el uso de una base de datos de entrenamiento, validación y test? ¿Cuál es la diferencia entre estas?

Es necesario el uso de estas diferentes bases de datos para desarrollar y evaluar modelos de aprendizaje automático de manera efectiva. Cada base de datos tiene un objetivo diferente por lo que se tienen que utilizar las tres para que el modelo sea capaz de generalizar bien a datos nuevos y no solo memorice los datos de entrenamiento. La base de datos de entrenamiento es la porción más grande de datos y se utiliza para entrenar el modelo. Así, el modelo se ajusta a esta información, aprendiendo patrones y características de las imágenes proporcionadas. Es importante que estos datos sean lo más variados y representativos posible. Por otro lado, la base de datos de validación se utiliza para realizar una comparación de diferentes algoritmos que fueron entrenados con la misma base de datos de entrenamiento. Por último, la base de prueba se utiliza al final del proceso de desarrollo del modelo para realizar una evaluación cuantitativa final de este. Estos datos son completamente independientes de las otras dos bases de datos, por lo que se realiza una evaluación crítica del modelo y proporciona una medida objetiva de como el modelo se comportará en el mundo real con datos nuevos.

1.2. ¿Cree que su base de datos está desbalanceada? Recuerde que queremos detectar glóbulos blancos

La base de datos proporcionada está desbalanceada ya que al comparar la cantidad de datos que hay de cada clase se evidencia que hay una mayor cantidad de datos de categoría 2 que el resto de clases.

1.3. ¿Qué implica que hayan imágenes sin nuestra clase de interés en entrenamiento?¿Cuál puede ser su implicación en test y validación?

Incluir imágenes que no pertenecen a nuestra clase de interés, en este caso, los glóbulos blancos, en la base de datos de entrenamiento puede diluir el enfoque del modelo y generar un sesgo hacia la clase mayoritaria (las imágenes sin glóbulos blancos). Esto puede ocasionar dificultades en la detección de la clase de interés debido a su representación más limitada en los datos de entrenamiento. Como resultado, el modelo podría aprender características irrelevantes de las imágenes sin glóbulos blancos, lo que, a su vez, podría afectar negativamente su capacidad para detectar con precisión la clase de interés en imágenes reales. Además, en las evaluaciones de validación y prueba, esto podría conducir a una impresión inflada y errónea del rendimiento del modelo. Por lo tanto, es esencial garantizar un conjunto de entrenamiento equilibrado con una representación adecuada de ambas clases, lo que permitirá al modelo aprender características relevantes para una detección precisa de la clase de interés.

1.4. ¿Qué afectaría el solo tener una base de datos de entrenamiento?¿Cómo lo solucionaría?

En un modelo podría llegar a ser problemático tener solo una base datos. Esto dificulta la evaluación de la capacidad de generalización del modelo y la detección del sobreajuste. Para solucionar esto, sería importante dividir los datos en diferentes conjuntos de entrenamiento, validación y prueba. Esta división asegura una evaluación independiente y una mejor capacidad de adaptación del modelo a situaciones del mundo real.

1.5. Esquema de anotaciones

En la figura 1 se observa la organización del archivo *_annotations.coco.json* de las carpetas de entrenamiento, validación y test. Este archivo JSON es guardado como un diccionario en el programa, el cual contiene cinco parejas llave-valor:

1. 'info': su valor es un diccionario con la información del conjunto de datos utilizado en el proyecto.

2. 'licenses': su valor es una lista de diccionarios, donde cada diccionario tiene la información de una licencia

para las imágenes utilizadas en el proyecto. En este caso, solo se utilizó la licencia de MIT.

3. 'categories': su valor es una lista de diccionarios, donde cada diccionario tiene la información de una categoría. En nuestro caso, tenemos cuatro categorías: células, plaquetas, glóbulos rojos y glóbulos blancos.

4. 'images': su valor es una lista de diccionarios, donde cada diccionario tiene la información de una imagen. Entonces, para cada imagen se tiene su id, su licencia, el nombre del archivo, sus dimensiones y la fecha de captura.

5. 'annotations': su valor es una lista de diccionarios, donde cada diccionario tiene la información de una anotación. Entonces, para cada anotación se tiene su id, el id de la imagen a la que pertenece la anotación, el id de la categoría a la que pertenece la anotación, la información necesaria para generar la *bbox* de la anotación: la ubicación en la imagen y sus dimensiones, entre otros.

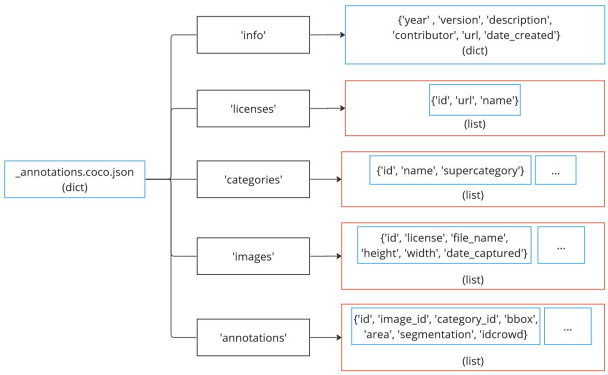

Figure 1. Esquema de la organización del archivo JSON de las anotaciones de cada carpeta. Los cuadros en azul corresponden a diccionarios, los cuadros en rojo a listas y las palabras en comillas simples corresponden a llaves de un diccionario.

El *image_id* es un identificador único que se asigna a cada imagen en un conjunto de datos. Es utilizado para diferenciar y asociar las anotaciones con imágenes específicas. Cada anotación de objeto en una imagen estará vinculada a su *image_id* correspondiente. El *category_id* es un identificador único asignado a cada categoría presente en un conjunto de datos. En nuestro caso, cada categoría representa un componente sanguíneo, como "célula", "plaqueta", "glóbulo rojo" y "glóbulo blanco". Las anotaciones se etiquetan con un *category_id* y un *image_id* para indicar qué tipo de componente está representado en qué imagen del conjunto de datos.

2. Visualización de imágenes

2.1. ¿Los canales RGB que tenemos son suficientes para diferenciar las células blancas?

Sí, ya que, el contraste existente entre lo que es una célula blanca y lo que no, es suficiente para hacer una detección de manera efectiva y con poco margen de error.

2.2. ¿Que busca resolver un problema de detección? ¿Cuáles son las anotaciones que usted tiene y cuáles son las predicciones que espera obtener?

Un problema de detección busca localizar en una imagen todas las instancias de un objeto de interés, en pocas palabras, localizar un objeto específico en una imagen. En esta ocasión, las anotaciones son un recuadro que encierra la totalidad del objeto de interés, es decir, el glóbulo blanco. En este orden de ideas, las predicciones que se desea obtener es que si yo tengo una anotación correspondiente a glóbulo blanco, la predicción a obtener sea un glóbulo blanco.

2.3. Proponga una metodología para determinar un buen umbral para la segmentación de los glóbulos blancos.

En primer lugar, utilizar la imagen en un canal rojo, ya que, este permite tener un mayor contraste entre el glóbulo blanco y todo lo demás. Posteriormente, teniendo en cuenta el nivel de gris que toman los píxeles pertenecientes al glóbulo blanco, seleccionar un k algo menor a ese nivel de gris, para que de esta manera, se oscurezcan todos los píxeles diferentes al glóbulo blanco y se aclaren todos los pertenecientes a un glóbulo blanco.

3. Resultados

En la figura 2 se observan dos imágenes aleatorias de la base de datos.

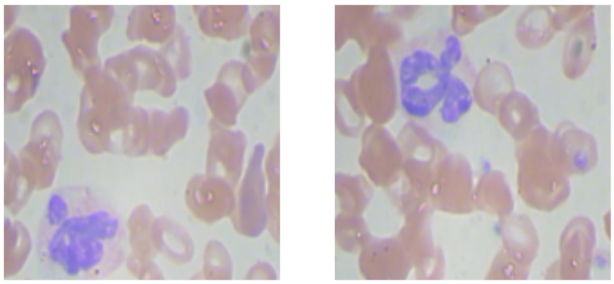

Figure 2. Imágenes de glóbulos rojos y blancos

En la figura 3 se observan dos columnas de imágenes donde la columna de la izquierda hace referencia a 4 imágenes originales de la base de datos proporcionada. Por

otro lado, se observa en la columna de la derecha las mismas imágenes pero con las anotaciones que encierran los glóbulos blancos. Se observa que para las 4 imágenes escogidas, se anotó de manera correcta la clase de interés, es decir, los glóbulos blancos.

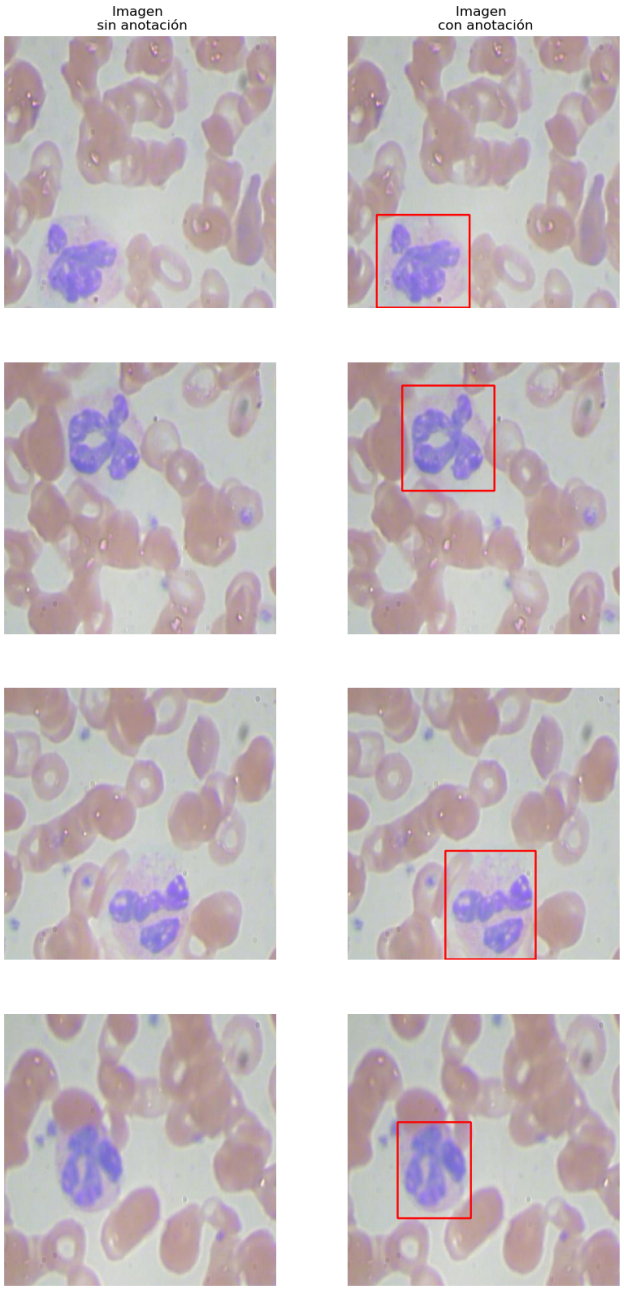

Figure 3. Imágenes de glóbulos blancos anotados

En la figura 4 se puede observar que se logró obtener las 4 imágenes deseadas por canales rojo, verde y azul respectivamente.

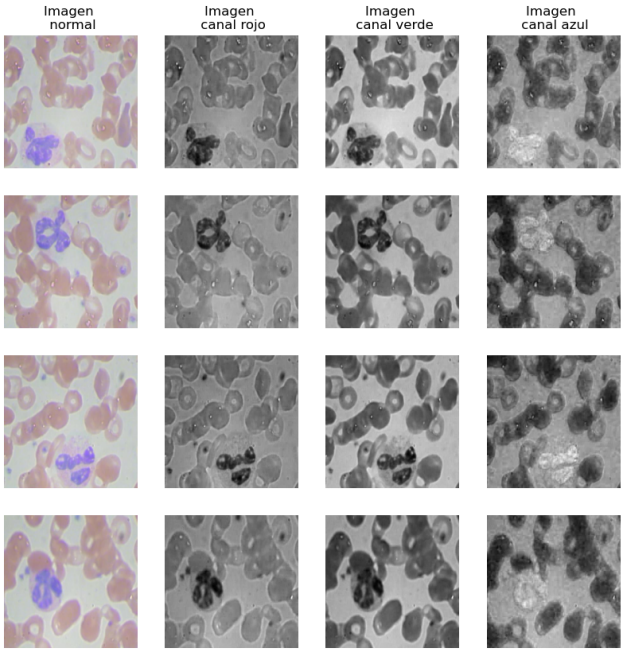

Figure 4. Imágenes de glóbulos blancos y rojos en canales rojo, verde y azul

En la figura 5 se puede observar que se logró umbralizar la imagen de manera correcta. Lo anterior debido a que para todos los píxeles con niveles de grises menores al umbral k seleccionado, se remplazaron por blancos y, todos aquellos mayores al umbral k, pasaron a ser negros.

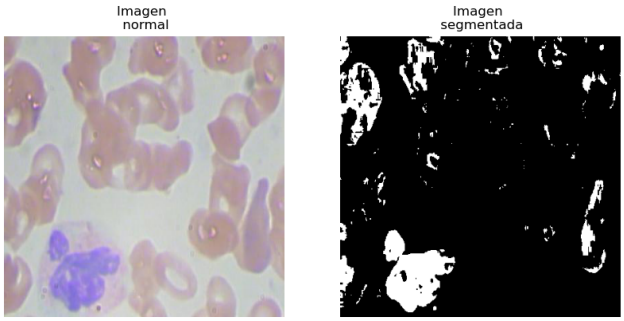

Figure 5. Umbralización de imagen con glóbulos blancos y rojos

Realizado en LaTeX.