Contexto de Aplicación
Este proyecto de machine learning surge en el campo de la microbiología, específicamente en el estudio de cianobacterias realizado por el grupo GAIA. Inicialmente, el desafío consistía en desarrollar una aplicación capaz de contar y clasificar cianobacterias a partir de imágenes microscópicas. Este problema complejo requería un enfoque que abordara tanto la segmentación y localización de los microorganismos como su posterior clasificación.

En la primera fase del proyecto, se centró en la segmentación y conteo de microorganismos. Se utilizó el modelo Faster SAM (Segment Anything Model) de Meta, que demostró una capacidad casi perfecta para segmentar y localizar microorganismos en las imágenes con guía humana. Esto representó un avance significativo en la automatización del proceso de conteo, reduciendo considerablemente el tiempo necesario en comparación con el conteo manual.

Con la segmentación y conteo resueltos, el siguiente paso es añadir la funcionalidad de clasificación a la aplicación. Este es el enfoque del presente proyecto. La idea es tomar las imágenes segmentadas por Faster SAM, extraer cada uno de los polígonos que representan microorganismos individuales y pasarlos por un clasificador basado en redes neuronales convolucionales utilizando TensorFlow.

Aunque el proyecto original se centra en cianobacterias, debido a acuerdos de confidencialidad con el grupo GAIA, para este curso se trabajará con datasets públicos relacionados con células sanguíneas como sustituto. Esto permitirá mantener la esencia del problema de clasificación de microorganismos, al tiempo que se trabaja con datos que pueden ser discutidos y compartidos abiertamente.

Objetivo de Machine Learning
El objetivo principal de este proyecto es desarrollar un modelo de clasificación robusto y eficiente capaz de identificar y categorizar diferentes tipos de microorganismos en imágenes microscópicas. Específicamente, buscamos:

- **Predecir el taxón o la clase** a la que pertenece cada microorganismo individual en una imagen previamente segmentada.
- **Lograr una alta precisión** en la clasificación, incluso en presencia de variabilidad morfológica dentro de las clases.

Para lograr estos objetivos, se utilizarán técnicas de aprendizaje profundo, específicamente redes neuronales convolucionales (CNN), que han demostrado ser efectivas en problemas de clasificación de imágenes. El modelo se entrenará utilizando datasets públicos de células sanguíneas, que servirán como sustituto para el estudio de cianobacterias.

Dataset

Se utilizarán cuatro datasets públicos relacionados con células sanguíneas y malaria como sustitutos del estudio de cianobacterias. A continuación, se presentan los datasets seleccionados:

1. **Blood Cells Image Dataset**
    - **Descripción:** Contiene 17,092 imágenes de células sanguíneas individuales normales.
    - **Clases:** 8 grupos (neutrófilos, eosinófilos, basófilos, linfocitos, monocitos, granulocitos inmaduros, eritroblastos y plaquetas).
    - **Enlace al dataset:** Blood Cells Image Dataset
2. **Blood Cell Images Dataset**
    - **Descripción:** Contiene 12,500 imágenes aumentadas de células sanguíneas.
    - **Clases:** 4 tipos (Eosinófilo, Linfocito, Monocito y Neutrófilo).
    - **Imágenes:** Aproximadamente 3,000 imágenes por clase, además de 410 imágenes originales.
    - **Enlace al dataset:** Blood Cell Images
3. **Blood Cell Detection Dataset**
    - **Descripción:** Contiene 364 imágenes de células sanguíneas con 4,888 etiquetas en tres clases: Glóbulos blancos (WBC), Glóbulos rojos (RBC) y Plaquetas.
    - **Formato:** Datos preprocesados para YOLOv5.
    - **Enlace al dataset:** Blood Cell Detection Dataset
4. **Malaria Bounding Boxes Dataset**
    - **Descripción:** Incluye 1,364 imágenes (~80,000 células) con células no infectadas (RBC y leucocitos) y células infectadas (gametocitos, anillos, trofozoítos y esquizontes).
    - **Etiquetas:** Incluye clases y coordenadas de cuadros delimitadores.
    - **Enlace al dataset:** Malaria Bounding Boxes Dataset

El tamaño total de los cuatro datasets descargados y comprimidos es de aproximadamente 4.58 GB.

Métricas de Desempeño

Para evaluar el desempeño del modelo de clasificación, se utilizarán las siguientes métricas:

1. **Precisión (Accuracy)**: Mide el porcentaje de predicciones correctas.
2. **Precisión por clase (Class-wise Precision)**: Indica el rendimiento del modelo para cada clase.
3. **Sensibilidad (Recall)**: Evalúa la capacidad del modelo para identificar correctamente todos los ejemplos de una clase.
4. **F1-score**: Combina la precisión y la sensibilidad en una única métrica armónica.
5. **Matriz de confusión**: Proporciona una visión más detallada de las predicciones correctas e incorrectas por clase.

Estas métricas ofrecerán una visión integral del rendimiento del modelo, permitiendo identificar posibles áreas de mejora.

Referencias y Resultados Previos

1. Faster SAM: Segment Anything Model. Meta AI.
2. Blood Cells Image Dataset. Kaggle.
3. Blood Cell Images. Kaggle.
4. Blood Cell Detection Dataset. Kaggle.
5. Malaria Bounding Boxes Dataset. Kaggle.

Conclusiones

El desarrollo de un modelo de clasificación robusto para microorganismos en imágenes microscópicas es un desafío alcanzable. La combinación de técnicas de aprendizaje profundo y el uso de datasets públicos de células sanguíneas como sustituto para el estudio de cianobacterias permitirá desarrollar un modelo capaz de manejar diferentes tipos de estructuras biológicas. El modelo se implementará utilizando TensorFlow y se evaluará a través de métricas de desempeño como la precisión, la precisión por clase, la sensibilidad y el F1-score.