Universidad del Valle de Guatemala
Facultad de ingeniería

Resultados iniciales

Clasificación de los orígenes de un coágulo de sangre en un accidente cerebrovascular.

Elisa Samayoa

Julio Avila 20333

Cayetano Molina 20211

Juan Fernando Ramirez

Javier Aguilar, 20611

Guatemala 29 de octubre del 2023

Introducción

Este proyecto tiene como objetivo principal la clasificación de los orígenes de los coágulos sanguíneos en AIS utilizando imágenes completas de patología digital. Por lo que se centra en diferenciar entre los dos principales subtipos de etiología del AIS: aterosclerosis cardíaca y aterosclerosis de la arteria grande. Al hacerlo, esperamos facilitar la prescripción de un tratamiento terapéutico post-AIS más efectivo y reducir la probabilidad de un segundo AIS.

Dado que un accidente cerebrovascular es la segunda causa de muerte a nivel mundial, con más de 700,000 personas en los Estados Unidos experimentando un accidente cerebrovascular isquémico cada año. Este proyecto busca abordar este problema crítico de salud pública mediante el desarrollo de un modelo de redes neuronales (CNN) que pueda identificar con precisión la etiología de los accidentes cerebrovasculares isquémicos agudos (AIS), lo que podría mejorar significativamente el tratamiento y la supervivencia del paciente.

Para lograr esto, crearemos un conjunto de algoritmos de procesamiento de imágenes para extraer y caracterizar las características patológicas relevantes en las imágenes completas de patología digital. Luego, entrenaremos y evaluaremos un modelo CNN utilizando este conjunto de datos, con el objetivo de alcanzar una precisión de por lo menos el 50% en la clasificación de la etiología de los coágulos sanguíneos en AIS.

Este proyecto se llevará a cabo con la base de datos extraidos de kaggle en colaboración con Mayo Clinic y su Registro de Tromboembolismo por Accidentes Cerebrovasculares (STRIP), que tiene como objetivo caracterizar histopatológicamente los tromboembolios de diversas etiologías. Creemos que este proyecto tiene el potencial para hacer una contribución significativa a nuestra comprensión y tratamiento del AIS, mejorando así las perspectivas para los pacientes afectados por esta devastadora condición.

Objetivos

Objetivo General

Desarrollar un modelo de inteligencia artificial efectivo y preciso que pueda distinguir entre las dos principales etiologías de los accidentes cerebrovasculares isquémicos agudos (AIS) - causa cardíaca y aterosclerosis de arteria grande - utilizando imágenes completas de patología digital, con el propósito de mejorar la identificación de los orígenes de los coágulos sanguíneos en accidentes cerebrovasculares mortales y facilitar la prescripción de un tratamiento terapéutico post-AIS más efectivo, reduciendo así la probabilidad de un segundo AIS.

Objetivos Específicos

1. Crear un conjunto de algoritmos de procesamiento de imágenes que permitan la extracción y caracterización de las características patológicas relevantes en las imágenes completas de patología digital relacionadas con los coágulos sanguíneos en AIS de origen cardíaco y aterosclerosis de arteria grande
2. Entrenar y evaluar un modelo de inteligencia artificial utilizando el conjunto de datos de imágenes completas de patología digital, con el objetivo de alcanzar una precisión de por lo menos el 50% en la clasificación de la etiología de los coágulos sanguíneos en AIS, diferenciando con alta sensibilidad y especificidad entre las causas cardíacas y la aterosclerosis de arteria grande.

Marco Teórico

Para efectos de la presente investigación de dicho proyecto, las variables de estudio y su relación entre cada una seran las siguientes: accidente cerebrovascular isquémico agudo (AIS), aterosclerosis cerebral, aterosclerosis de la arteria grande, coágulos sanguíneos, imágenes completas de patología digital, algoritmos, algoritmos de procesamiento de imágenes, algoritmo CNN e inteligencia artificial.

El accidente cerebrovascular isquémico agudo (AIS) es una condición médica que ocurre cuando un coágulo de sangre bloquea el flujo sanguíneo a una parte del cerebro, provocando daño o muerte de las células cerebrales. El AIS es la segunda causa de muerte a nivel mundial y una de las principales causas de discapacidad. Según la Organización Mundial de la Salud, se estima que 15 millones de personas sufren un accidente cerebrovascular cada año, de los cuales 5 millones mueren y otros 5 millones quedan con discapacidad permanente. El AIS tiene un gran impacto en la calidad de vida de los pacientes y sus familias, así como en los sistemas de salud y sociales. Además, según Medilineplus existen 2 tipos de AIS Isquémico y hemorrágico. El ataque cerebral isquémico es el tipo más común. En general, es causado por un coágulo sanguíneo que bloquea o tapa un vaso sanguíneo en el cerebro. Esto evita que la sangre fluya hacia éste órgano. En cuestión de minutos, las células del cerebro comienzan a morir. Otra causa es la estenosis o estrechamiento arterial. Esto puede suceder debido a la aterosclerosis, una enfermedad en la que se acumula placa en las arterias. Los ataques isquémicos transitorios se producen cuando la sangre no llega al cerebro por unos instantes. Tener un ataque isquémico transitorio puede significar que usted está en riesgo de sufrir un derrame cerebral más grave (Medilineplus, 2023).

La aterosclerosis cerebral llega a ser una enfermedad que afecta a las arterias cerebrales, que son las que llevan sangre al cerebro. La aterosclerosis cerebral se produce cuando se forman placas de ateroma en las paredes de las arterias, que pueden estrechar o bloquear el paso de la sangre y causar isquemia, infarto o hemorragia cerebral. La aterosclerosis cerebral es una de las principales causas de accidentes cerebrovasculares, demencia y deterioro cognitivo. Además segun Medlineplus La arterioesclerosis es una afección en la cual placa se acumula dentro de las arterias. Placa es una sustancia pegajosa compuesta de grasa, colesterol, calcio y otras sustancias que se encuentran en la sangre. Con el tiempo, esta placa se endurece y angosta las arterias. Eso limita el flujo de sangre rica en oxígeno (Medilineplus, 2023).

La Aterosclerosis de la arteria grande llega a ser un tipo especifico de aterosclerosis que afecta a las arterias de gran calibre, como la aorta, la carótida o la femoral. La aterosclerosis de la arteria grande se produce cuando se forman placas de ateroma en las paredes de las arterias, que pueden romperse y liberar fragmentos que viajan por el torrente sanguíneo y pueden obstruir otras arterias más pequeñas, causando accidentes cerebrovasculares, isquemia o aneurismas.

Los coágulos sanguíneos son masas sólidas que se forman cuando la sangre se coagula, es decir, cuando se activan los mecanismos de defensa del organismo para detener una hemorragia. Los coágulos sanguíneos pueden ser beneficiosos cuando se producen en una herida, pero pueden ser peligrosos cuando se forman en el interior de los vasos sanguíneos, debido a esto pueden llegar a bloquear el flujo de sangre y causar trombosis, embolia o accidentes cerebrovasculares. Además según Medlineplus los coágulos de sangre se pueden formar o viajar a los vasos sanguíneos de las extremidades, los pulmones, el cerebro, el corazón y los riñones (Medilineplus, 2023).

Las imágenes completas de patología digital son imágenes digitales de alta resolución que capturan la totalidad de una muestra histológica, es decir, un tejido extraído del cuerpo humano y procesado para su análisis microscópico. Las imágenes completas de patología digital permiten visualizar y estudiar las características morfológicas y moleculares de los tejidos, así como realizar diagnósticos, pronósticos y tratamientos personalizados. Además según Leica Biosystems la patología digital se pueden aplicar en herramientas automatizadas de análisis de imágenes que ayuden en la interpretación y cuantificación de la expresión de biomarcadores dentro de secciones de tejido (Leica,2023).

El origen de toda secuencia de pasos de cualquier modelo matemático, estadístico, computarizado o redes neuronales se guían a base de algoritmos los cuales son secuencias finitas y ordenadas de pasos o instrucciones que permiten resolver un problema o realizar una tarea. Los algoritmos se pueden implementar en programas informáticos que ejecutan las instrucciones de forma automática y eficiente. Los algoritmos se pueden clasificar según su complejidad, su diseño, su aplicación o su tipo de datos.

Los algoritmos de procesamiento de imágenes: Son algoritmos que manipulan o transforman imágenes digitales con el fin de mejorar su calidad, extraer información relevante, detectar objetos o patrones, segmentar regiones o aplicar efectos. Los algoritmos de procesamiento de imágenes se basan en operaciones matemáticas y estadísticas que se aplican a los píxeles que componen las imágenes. Según academiaLab esto también nos permite aplicar una gama mucho más amplia de algoritmos a los datos de entrada y puede evitar problemas como la acumulación de ruido y distorsión durante el procesamiento (academiaLab, 2023).

Un Algoritmo CNN es un tipo de algoritmo de aprendizaje profundo que utiliza redes neuronales convolucionales (CNN) para procesar imágenes. Las redes neuronales convolucionales son modelos computacionales inspirados en el funcionamiento del cerebro humano, que consisten en capas de neuronas artificiales que se conectan entre sí y aprenden a partir de los datos. Las redes neuronales convolucionales utilizan filtros o kernels que se deslizan sobre las imágenes y extraen características visuales jerárquicas y abstractas. Además, se escogió este algoritmo de reconocimiento de objetos en imágenes debido a que es uno de los mas reconocidos y efectivos para el proyecto e investigación que estamos realizando dado que “se encargan de extraer características relevantes de la imagen, y capas completamente conectadas, que realizan la clasificación final” (InteligenciaArificial.Science, 2023).

La inteligencia artificial vino a cambiar cómo nos desarrollamos de manera más efectiva todas nuestras actividades diarias dado que es la disciplina científica que estudia cómo crear sistemas informáticos capaces de realizar tareas que normalmente requieren inteligencia humana, como percibir, razonar, aprender, comunicarse o tomar decisiones. La inteligencia artificial se puede dividir en dos ramas: la inteligencia artificial débil o aplicada, que se enfoca en resolver problemas específicos; y la inteligencia artificial fuerte o general, que busca emular la inteligencia humana en todos sus aspectos. Asimismo, según Tableau, uno de los softwares más utilizados para todo el análisis y visualización de datos en el mundo, la inteligencia artificial mediante la creación de algoritmos y sistemas especializados, las máquinas pueden llevar a cabo procesos propios de la inteligencia humana, como aprender, razonar o autocorregirse (Tableau, 2023).

Estos términos y conceptos anteriormente expuestos se relacionan entre sí para poder lograr el objetivo del proyecto, los cuales se llegan a relacionar de la siguiente manera:

El proyecto buscó crear un conjunto de algoritmos de procesamiento de imágenes que permitiera la extracción y caracterización de las características patológicas relevantes en las imágenes completas de patología digital relacionadas con los coágulos sanguíneos en AIS (accidentes cerebrovasculares isquémicos agudos) de origen cardiaco y aterosclerosis de arteria grande. Para ello, utilizamos algoritmos de aprendizaje profundo, en particular algoritmos CNN, que son capaces de procesar imágenes de alta resolución y extraer características visuales complejas y discriminativas. Además, se desarrolló un modelo de inteligencia artificial con redes neuronales efectivo y preciso que pueda distinguir entre las dos principales etiologías de los AIS: causa cardíaca y aterosclerosis de arteria grande, utilizando las características extraídas de las imágenes completas de patología digital. Por lo que con este modelo, se pudo mejorar la identificación de los orígenes de los coágulos sanguíneos en accidentes cerebrovasculares mortales y facilitar la prescripción de un tratamiento terapéutico post-AIS más efectivo, reduciendo así la probabilidad de un segundo AIS.

Metodología

El proyecto se realizó en el lenguaje Python, a través del API de Kaggle, debido a que el conjunto de datos utilizado almacena alrededor de 395GB y no era eficiente descargarla localmente en las computadoras de los miembros del equipo. En Kaggle, se utilizó una RAM de 30 GB y en CPU fueron 390.

Primero, se importaron bibliotecas como NumPy, pandas, y varias bibliotecas relacionadas con el procesamiento de imágenes y el aprendizaje automático, tales como Keras de Tensor Flow, ya que es muy útil para desarrollar redes neuronales como la que se utilizó, también se utilizó la librería sklearn, que tuvo gran importancia en el desarrollo del proyecto.

Posteriormente, se procedió a la carga de los archivos CSV: train, test y other, el cual contenía otro tipo de enfermedades no etiquetadas, y con el análisis exploratorio de datos, el cual incluyó la verificación de la forma de los conjuntos de datos, la visualización de las primeras filas, y la comprobación de la información sobre las columnas, como la cantidad de datos faltantes. Después, se realizó el procesamiento de datos y creación de rutas de imágenes creando una nueva columna en el conjunto de datos de entrenamiento que contiene rutas de imágenes basadas en identificadores de imágenes y realizando comprobaciones sobre datos faltantes y la cantidad de clases únicas en el conjunto de entrenamiento.

Al finalizar, se generan gráficos y visualizaciones para comprender la distribución de las etiquetas y otras características en los datos de imágenes. Como las imágenes tenían dimensiones muy grandes, la mayoría con muchos espacios en “blanco”, se preprocesaron para reducir su resolución y calcular la proporción de área blanca en cada imagen. Luego, se calcularon los pesos de clase para abordar el desequilibrio en las etiquetas y se utilizaron en el entrenamiento del modelo.

Kaggle ya proporcionaba la división de train y test de las imágenes, por lo que no fue necesario realizar una separación adicional. Para el train, se utilizaron 754 imágenes, mientras que para el test fueron 280. Con esto, se creó un modelo de aprendizaje automático basado en la arquitectura ResNet50. El modelo se compila con la función de pérdida, el optimizador y las métricas necesarias.

Más adelante, se aplicó un aumento de datos en el conjunto de entrenamiento, lo que implicó realizar transformaciones en las imágenes para aumentar la cantidad de datos de entrenamiento y mejorar la capacidad del modelo para generalizar. Luego, se entrenó el modelo utilizando el conjunto de datos de entrenamiento con una serie de épocas con los generadores de imágenes creados previamente.

Por último, se generaron gráficos para visualizar la precisión, la pérdida y el puntaje F1 durante el entrenamiento del modelo. Una vez se alcanzó el nivel mínimo esperado en las mediciones del accuracy, f1 score y demás, se procedió a guardar el modelo entrenado en un archivo y se procesaron y predijeron nuevos resultados con base al modelo entrenado.

Finalmente, se mostraron las predicciones junto con la probabilidad de clase para cada imagen y prueba y se procedió a elaborar la aplicación como interfaz gráfica para los resultados.

El algoritmo seleccionado fue una red neuronal convolucional (CNN) debido a que tienen muchas ventajas que eran muy útiles para el proyecto en cuestión. La principal ventaja que nos hizo

seleccionar este tipo de red es que tienen gran capacidad de aprender automáticamente características complejas de las imágenes, lo que elimina la necesidad de extraer manualmente dichas características. Además, estas son efectivas para identificar patrones en las imágenes independientemente de su ubicación en la imagen y usan menos parámetros que las redes neuronales totalmente conectadas, haciendo las CNN más eficientes.

Resultados y Análisis de Resultados

El proceso de preparación de datos para el análisis y modelado se dividió en varias etapas. En primer lugar, se llevó a cabo la carga y exploración de datos. Los conjuntos de datos se leyeron para comprender la estructura y características de los datos. Esto incluyó una revisión inicial de las primeras filas, información de columnas, tipos de datos, y la identificación de posibles valores nulos o faltantes mediante funciones como head(), info() y value_counts().

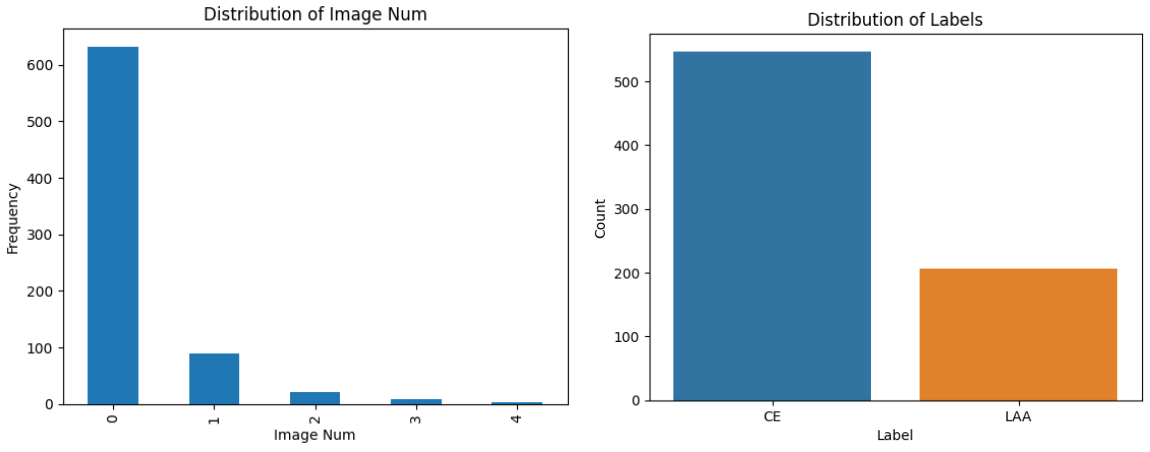
```
Train_df
     image_id  center_id patient_id  image_num label
0    006388_0         11     006388          0    CE
1    008e5c_0         11     008e5c          0    CE
2    00c058_0         11     00c058          0   LAA
3    01adc5_0         11     01adc5          0   LAA
4    026c97_0          4     026c97          0    CE
..        ...        ...        ...        ...   ...
749  fe9645_0          3     fe9645          0    CE
750  fe9bec_0          4     fe9bec          0   LAA
751  ff14e0_0          6     ff14e0          0    CE
752  ffec5c_0          7     ffec5c          0   LAA
753  ffec5c_1          7     ffec5c          1   LAA

[754 rows x 5 columns]


Test_df
   image_id  center_id patient_id  image_num
0  006388_0         11     006388          0
1  008e5c_0         11     008e5c          0
2  00c058_0         11     00c058          0
3  01adc5_0         11     01adc5          0


Other_df
     image_id patient_id  image_num  other_specified    label
0    01f2b3_0     01f2b3          0              NaN  Unknown
1    01f2b3_1     01f2b3          1              NaN  Unknown
2    02ebd5_0     02ebd5          0              NaN  Unknown
3    0412ab_0     0412ab          0              NaN  Unknown
4    04414e_0     04414e          0  Hypercoagulable    Other
..        ...        ...        ...              ...      ...
391  faaa7e_0     faaa7e          0              NaN  Unknown
392  fd0f11_0     fd0f11          0              NaN  Unknown
393  fd0f11_1     fd0f11          1              NaN  Unknown
394  fd83c3_0     fd83c3          0              NaN  Unknown
395  febb2b_0     febb2b          0              NaN  Unknown
```

Antes de preprocesar las imágenes, se realizaron gráficas para comprender el comportamiento de los datasets proporcionados.



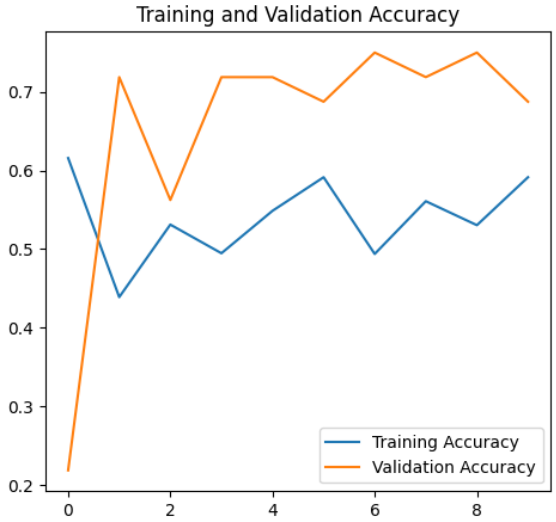
Como pudo observarse, con los datos originales, la mayoría de los pacientes tan solo tenían 1 imagen (o ninguna) para analizar, de las cuales, en su mayoría son CE, es decir, que están etiquetadas como infarto cerebral cardioembólico, mientras que menos de la mitad es arterioesclerosis. Esto implicó el desbalance en el dataset, lo que llevó a balancear el train y dejar desvalanceado el test.

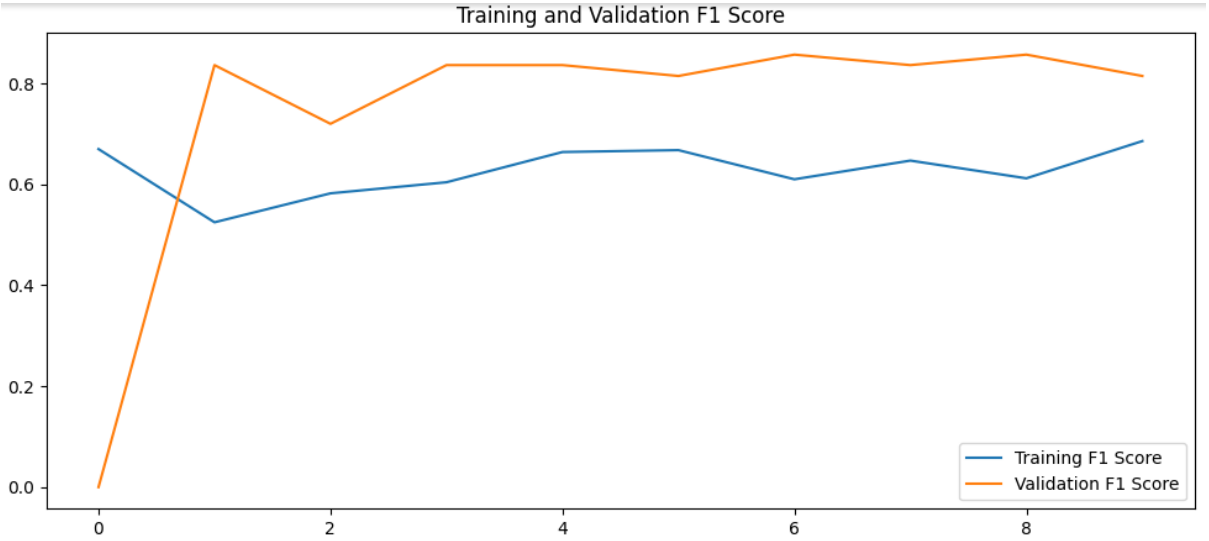
Posteriormente, se aplicó un proceso de procesamiento específico para las imágenes. Se implementó una función para reducir la resolución de las imágenes, permitiendo así trabajar con escalas más manejables. Además, se extrajeron características de las imágenes, lo que contribuyó a identificar áreas de interés dentro de las imágenes.
En cuanto a la preparación de los datos para el modelado, se generaron etiquetas a partir de la información disponible. Se procedió a crear conjuntos de datos de entrenamiento y pruebas, normalizándolos y dividiéndolos de manera adecuada para ser utilizados en el entrenamiento del modelo.

Durante el proceso de entrenamiento, se empleó un modelo pre-entrenado, en este caso ResNet50, y se ajustaron las capas superiores del modelo para adaptarlas al problema. Los modelos de clasificación fueron compilados y entrenados con métricas específicas, tales como el F1 Score, utilizando conjuntos de datos de entrenamiento y validación.
Para la evaluación se analizó la precisión y la pérdida durante el entrenamiento y la validación para comprender el rendimiento de los modelos. Finalmente, se aplicó el modelo a un conjunto de datos de prueba para obtener predicciones. Los resultados obtenidos fueron interpretados para evaluar la eficacia del modelo.

Durante el proceso de entrenamiento de los modelos, se utilizaron estrategias específicas para mejorar su rendimiento. Se optó por emplear modelos pre-entrenados como ResNet50, congelando algunas capas para retener patrones previos, y ajustando solo las capas superiores para adaptar el modelo al nuevo conjunto de datos. Además, se aplicaron técnicas de aumento de datos para mejorar la capacidad de generalización del modelo y se realizaron ajustes en los hiperparámetros, como la tasa de aprendizaje y el tamaño del lote, para influir en la precisión y convergencia del modelo.
Estas estrategias permitieron evaluar y validar el rendimiento del modelo en datos de entrenamiento y validación..

El modelo tiene resultados positivos y aún cuenta con margen de mejora si se entrena con una mayor cantidad de imágenes.
Tiene actualmente accuracy de 0.6875 y val_f1_score de 0.8148.





Este modelo lo comparamos con otro modelo que se realizó, también CNN con diferentes parámetros, el cual tuvo accuracy de 0.675 y val_f1_score de 0.8036

```
5/5 [==============================] - 0s 9ms/step - loss: 0.2460 - rmse: 0.4958 - accuracy: 0.6755 - binary_ac
curacy: 0.6755 - f1_score: 0.8036
```

El primer modelo demostró tener mejores resultados, por lo que fue el seleccionado para el desarrollo del proyecto.

En cuanto a la aplicación, se desarrolló una aplicación que permite al usuario ingresar alguna foto de un coágulo y después de un preprocesamiento, enseña el resultado de lo que se obtiene con los modelos utilizados.
Además, se puede cambiar el modelo que se quiere usar o bien, si se quieren usar todos. Con ello también se despliegan gráficas si es que el usuario lo desea, demostrando la eficiencia de los modelos que corrieron al ingresar la foto.

Bibliografía

- *Accidente cerebrovascular isquémico: MedlinePlus en español.* (n.d.). Retrieved October 30, 2023 from https://medlineplus.gov/spanish/ischemicstroke.html
- *Arterioesclerosis | Ateroesclerosis | MedlinePlus en español.* (n.d.). Retrieved October 30, 2023 from https://medlineplus.gov/spanish/atherosclerosis.html
- *Coágulos sanguíneos: MedlinePlus en español.* (n.d.). Retrieved October 30, 2023 from https://medlineplus.gov/spanish/bloodclots.html
- Heffner, S., Colgan, O., & Doolan, C. (n.d.). *¿Qué es la patología digital? | Leica Biosystems.* Leica Biosystems. Retrieved October 30, 2023 from https://www.leicabiosystems.com/es/knowledge-pathway/digital-pathology/
- *Procesamiento de imágenes _ AcademiaLab.* (n.d.). Retrieved October 30, 2023 from https://academia-lab.com/enciclopedia/procesamiento-de-imagenes/
- *Algoritmos clave para el reconocimiento de objetos en imágenes - Inteligencia Artificial.* (n.d.). Retrieved October 30, 2023 from https://inteligenciaartificial.science/algoritmos-clave-para-el-reconocimiento-de-objetos-en-imagenes/
- ¿Qué es la Inteligencia artificial? Definición, historia y aplicaciones. (n.d.). Tableau. https://www.tableau.com/es-mx/data-insights/ai/what-is.