ANTEPROYECTO TRABAJO FIN DE MÁSTER

Detección de tumores de cáncer de mama mediante técnicas de Computer Vision: Un enfoque para la mejora del diagnóstico temprano

Componentes:

María Inmaculada García Hernández

Álvaro Ladrón de Guevara Garcés

Ángel Martín Heras

Diego Muñoz Herranz

Tutores:

Teno González Dos Santos

David Sanz Díaz

Madrid, 4 de junio 2023

ÍNDICE

INTRODUCCIÓN

El cáncer de mama es el tumor maligno más frecuente entre la población femenina[1],[2]. La importancia de la detección temprana del cáncer de mama mediante el uso de la mamografía y otras técnicas es fundamental, ya que cambian el pronóstico de la enfermedad.

Los sistemas de inteligencia artificial pueden ayudar a la detección temprana de este tipo de tumores, en concreto trataremos el de tipo IDC (Invasive Ductal Carcinoma)[3], que es el tipo más común de cáncer de seno.

Este cáncer comienza en el revestimiento de los conductos galactóforos (conductos de la mama) los cuales son responsables de transportar la leche desde los lóbulos mamarios hasta el pezón durante la lactancia[1]. Estos conductos están revestidos por células epiteliales que normalmente se dividen y se renuevan de manera controlada. Sin embargo, en ocasiones, debido a diversos factores, estas células pueden experimentar cambios en su material genético, lo que desencadena un proceso de transformación maligna.

Cuando el cáncer se origina en los conductos galactóforos y se disemina fuera de ellos, se clasifica como un carcinoma ductal invasivo. Esto significa que las células cancerosas han atravesado la barrera del revestimiento del conducto y han invadido los tejidos circundantes, lo que les permite propagarse a otras áreas de la mama[4].

¿QUIÉNES SOMOS?

- **María Inmaculada García Hernández**

Licenciada en Físicas por la Universidad Complutense de Madrid con un año de ERASMUS en la universidad de Lund en Suecia.
Doctora Ingeniera Aeronáutica por la Universidad Politécnica de Madrid.

- **Álvaro Ladrón de Guevara Garcés**

Graduado en Matemáticas por la Universidad de Cádiz y actualmente formando parte de una multinacional en el ámbito de Consultoría, realizando funciones de diseño e implementación de soluciones a medida para los procesos financieros de diversas empresas.

- **Ángel Martín Heras**

Graduado en Ingeniería Matemática con especialidad en Economatemática por la Universidad Complutense de Madrid y dos años de experiencia en el sector financiero como analista de datos.

Diseñó un modelo eficiente de evaluación crediticia usando el mínimo de variables socioeconómicas posibles para un caso con datos reales como Trabajo de Fin de Grado. A partir de ahí continúa formándose en Inteligencia Artificial y Machine Learning**.**

- **Diego Muñoz Herranz**

Graduado en Desarrollo de Aplicaciones Multiplataforma por la Escuela Superior De Formación Audiovisual, Animación 3D y Nuevas Tecnologías y dos años de experiencia en el mundo de los datos, realizando en primera instancia análisis financiero para una entidad bancaria y actualmente realizando análisis y procesamiento de datos en la parte de auditoría en una multinacional.

Junto a un equipo de trabajo diseñó una alternativa a la aplicación Playtomic, realizando una aplicación IOS y un sistema de aperturas automáticas de puertas aplicando técnicas de IOT con Python.

ESTADO ACTUAL DE LA INVESTIGACIÓN

El uso de machine Learning en el estudio de imágenes de cáncer de mama se está volviendo cada vez más frecuente, ofreciendo nuevas oportunidades para la detección temprana y el análisis preciso de la enfermedad.

En general, para cualquier algoritmo que se quiera implementar para esta tarea se utiliza el enfoque CLAHE[5],[6] para mejorar la calidad de las imágenes y eliminar el posible ruido.

Adicionalmente, el rango de algoritmos que se usan para la investigación de este campo es muy grande, usando modelos de SVM, Random Forest y redes neuronales, aplicados tras identificar y delimitar las regiones de interés en las imágenes con las que se van a entrenar dichos modelos.

En el campo del estudio de imágenes de cáncer de mama, los avances en el uso de machine learning han permitido no solo el desarrollo de modelos para la detección temprana de la enfermedad, sino también para determinar la etapa en la que se encuentra. Además, estos modelos han demostrado ser eficaces en evaluar la efectividad del tratamiento[7]. Mediante el análisis de imágenes médicas, se pueden extraer características relevantes y utilizar algoritmos de aprendizaje automático para realizar una evaluación precisa de la progresión del cáncer y para determinar la respuesta al tratamiento. Esta capacidad de los modelos de machine learning brinda

nuevas oportunidades para un diagnóstico más preciso y una planificación de tratamiento personalizada en la lucha contra el cáncer de mama.

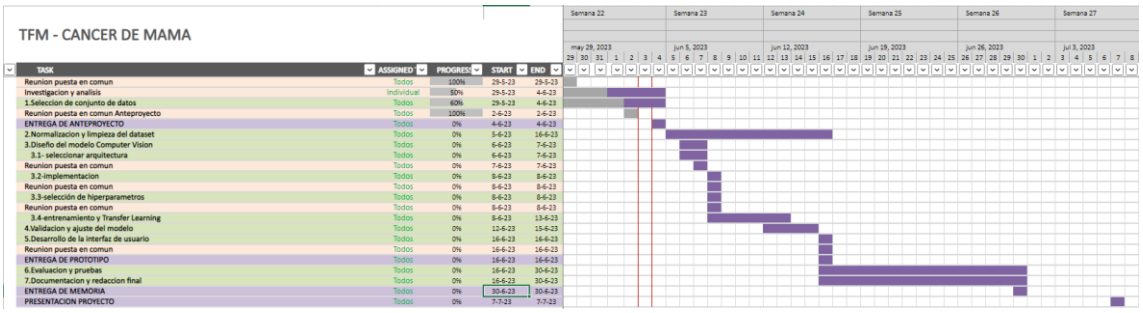
OBJETIVOS

En este epígrafe, se expondrán los objetivos que se pretenden alcanzar con la ejecución de este proyecto. El principal de ellos es desarrollar un modelo capaz de identificar si las masas tumorales son benignas o malignas. Para alcanzar esta meta, es necesaria la consecución de una serie de objetivos específicos. Estos son:

- Recopilar y preparar un dataset de imágenes médicas etiquetadas, el cual contenga tanto casos positivos como negativos de cáncer de mama.
- Utilizar modelos preentrenados utilizando técnicas de aprendizaje por transferencia que nos permitan reducir el volumen de datos del dataset inicial.
- Realizar un análisis exploratorio sobre los datos para lograr una comprensión óptima de sus características.
- Construir un modelo de red neuronal convolucional (CNN) capaz de detectar cáncer de mama en imágenes médicas. Este debe ser capaz de extraer propiedades relevantes de las imágenes.
- Aplicar técnicas de segmentación semántica para obtener información más detallada y precisa sobre el tamaño y ubicación del tumor.
- Entrenar el modelo haciendo uso de los datos ya preparados. Ajustar pesos de la red convolucional y medir la precisión del algoritmo con el conjunto de datos de entrenamiento. Encontrar hiperparámetros adecuados para las imágenes.
- Ratificar el modelo mediante la aplicación de este sobre un conjunto de datos reservado para testeo.
- Optimizar el algoritmo utilizando técnicas de ajuste de hiperparámetros.
- Realizar pruebas exhaustivas sobre el sistema completo para garantizar la precisión, eficiencia y fiabilidad del modelo.
- Desarrollar una interfaz que presente los resultados de aplicar el modelo después de recibir imágenes como entrada por parte del usuario.
- Documentar los resultados obtenidos.

PLAN DE TRABAJO

Para la realización del plan de trabajo, hemos creado un cronograma en el que hemos dividido las tareas y el cuál se rellenará a lo largo del proceso de desarrollo. En este cronograma hemos definido los deadlines principales de entrega para poder realizar una gestión correcta de todo el proceso.



Incluimos enlace al Excel en la parte bibliográfica[8]

VIABILIDAD DE RECURSOS

Para la realización del proyecto planteado utilizaremos Python en su mayoría, debido a su gran capacidad para la realización de modelos de inteligencia artificial gracias a las librerías open source para este trabajo. Detallaremos 3 de las ideas para realizar la tarea de clasificación.

- Redes Neuronales Convolucionales (CNN): Las CNN son una de las técnicas más utilizadas en Computer Vision. Estas redes están diseñadas específicamente para procesar imágenes y han demostrado un gran resultado en la tarea de clasificación.
- Aprendizaje por Transferencia (Transfer Learning): esta técnica permite aprovechar el conocimiento aprendido por una red neuronal previamente entrenada en un conjunto de datos grandes y aplicarlo a una tarea específica. Este punto es muy interesante cuando tenemos un conjunto de datos limitado o tenemos recursos limitados de cálculo.
- Segmentación Semántica: esta técnica se centra en la asignación de etiquetas a píxeles individuales de una imagen. Para nuestro caso en concreto nos permitiría identificar y delimitar regiones interesantes de las imágenes, como masas o microcalcificaciones.

Para el desarrollo de estas tareas utilizaremos diferentes librerías que nos permitan realizar modelos de inteligencia artificial.

- TensorFlow: framework que utilizaremos para construir y entrenar redes neuronales profundas.
- PyTorch: este framework proporciona una interfaz flexible y fácil de usar para entrenar modelos.
- Keras: framework de aprendizaje profundo.
- OpenCV: es una biblioteca de visión por computadora y procesamiento de imágenes de código abierto. Proporciona una amplia gama de algoritmos y funciones para el procesamiento de imágenes y vídeos.

Aparte de estas librerías tenemos planificado usar otras comunes como Pandas, Numpy, SKLearn, con el objetivo de realizar un análisis exhaustivo del proyecto. Empezando por el uso de modelos de inteligencia artificial básicos y terminando con modelos más avanzados, incluyendo el uso de modelos de aprendizaje por transferencia.

INFORME DE RIESGOS

Para los proyectos de tecnología es importante la realización de un plan de riesgos que evalúe e identifique las características del proyecto y nos permita saber si es un proyecto viable o no. En caso positivo, identificar que debemos de tener en cuenta en la realización de estos.

Para el análisis de riesgos, en nuestro caso en particular, lo hemos dividido en 5 puntos.

1. Falta de datos
    a. Identificación del Riesgo: disponibilidad limitada de datos de mamografías etiquetados que pueden dificultar el entrenamiento y la evaluación del modelo.
    b. Mitigación: búsqueda de bases de datos públicas o la colaboración de instituciones médicas para obtener acceso a un conjunto de datos adecuado. Exploración de técnicas de aumento de datos para ampliar el conjunto.
2. Calidad de los datos
    a. Identificación del Riesgo: los datos de mamografías pueden contener ruido, artefactos o información incompleta, esto puede afectar a la precisión del modelo.
    b. Mitigación: realizar un preprocesamiento de los datos para limpiar el ruido y eliminar los artefactos, así como la realización de un control de calidad de los datos y verificar su integridad antes de utilizarlos en el modelo.

3. Selección de características inadecuadas
    a. Identificación del Riesgo: la elección incorrecta de características dentro de las imágenes puede llevar a un modelo ineficiente o con un rendimiento deficiente.
    b. Mitigación: investigación para realizar una selección correcta y concreta de las mejores características. También podemos realizar técnicas de extracción automática de características.
4. Sobreajuste del modelo (overfitting)
    a. Identificación del Riesgo: un modelo sobreajustado puede tener un rendimiento muy bueno en los datos de entrenamiento, pero deficiente en los datos de prueba o datos nuevos.
    b. Mitigación: utilización de técnicas de validación cruzada y división adecuada de los conjuntos de datos en entrenamiento, validación y prueba. Aplicación de técnicas de regularización, para evitar este sobreajuste.
5. Ética y privacidad
    a. Identificación del Riesgo: los datos médicos, como las mamografías, contienen información sensible que requiere consideraciones éticas y de privacidad adecuadas.
    b. Mitigación: cumplir con las regulaciones y directrices éticas relevantes en el manejo y almacenamiento de datos médicos. Anonimizar y encriptar los datos según sea necesario. Obtener el consentimiento informado de los pacientes o asegurarse de que los datos utilizados sean completamente anónimos.

RESULTADOS QUE PUEDE GENERAR EL PROYECTO

El uso de la visión por computadora para la detección y predicción del cáncer de mama puede generar una serie de resultados prometedores, mejorando tanto la precisión, como la eficiencia de los procesos de diagnóstico y tratamiento.

Algunos de estos resultados podrían ser:

- **Detección temprana**: El principal beneficio del uso de la visión por computadora en la detección del cáncer de mama es la capacidad de detectar la enfermedad en sus etapas más tempranas, lo que puede aumentar significativamente las tasas de supervivencia. Los algoritmos pueden ser entrenados para identificar anomalías y patrones sutiles en las imágenes de mamografías que pueden pasar desapercibidos por el ojo humano.
- **Mayor precisión**: La visión por computadora puede disminuir la tasa de falsos positivos y falsos negativos en el diagnóstico del cáncer de mama, reduciendo así el estrés emocional y los costos asociados con pruebas adicionales o tratamientos innecesarios.

- **Automatización del screening**: Los sistemas de visión por computadora pueden procesar un gran número de imágenes en un tiempo relativamente corto, lo que podría liberar a los radiólogos y otros profesionales médicos para centrarse en los casos más complejos y en el cuidado directo del paciente.
- **Evaluación del riesgo personalizado**: Con la ayuda de la visión por computadora y los algoritmos de aprendizaje automático, es posible desarrollar un modelo de evaluación del riesgo personalizado que tome en cuenta las características individuales de los pacientes.
- **Investigación y aprendizaje continuo**: Los datos y los resultados generados a través de este tipo de proyecto pueden ser utilizados para mejorar los algoritmos, permitiendo una mejora continua en la precisión y la eficacia de la detección y el tratamiento del cáncer de mama.

BIBLIOGRAFÍA

- [1] Diccionario de cáncer del NCI. (n.d.). Instituto Nacional Del Cáncer. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/carcinoma-ductal-invasivo\
- [2] Siegel, R. L., Miller, K. A., Fuchs, H. E., & Jemal, A. (2021). Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians, 71(1), 7–33. https://doi.org/10.3322/caac.21654
- [3] Invasive Ductal Carcinoma (IDC): Grade, Symptoms & Diagnosis. (n.d.). https://www.breastcancer.org/types/invasive-ductal-carcinoma
- [4] Harbeck, N., & Gnant, M. (2017). Breast cancer. The Lancet, 389(10074), 1134–1150. https://doi.org/10.1016/s0140-6736(16)31891-8
- [5] Chaudhury, S., Krishna, A. N., Gupta, S., Sankaran, K. S., Khan, S., Sau, K., Raghuvanshi, A., & Sammy, F. (2022). Effective Image Processing and Segmentation-Based Machine Learning Techniques for Diagnosis of Breast Cancer. Computational and Mathematical Methods in Medicine, 2022, Article ID 6841334. https://www.hindawi.com/journals/cmmm/2022/6841334/
- [6] Avcı, H., & Karakaya, J. (2023). A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning. Diagnostics, 13(3), 348 https://www.mdpi.com/2075-4418/13/3/348
- [7] Sammut, S. J., Crispin-Ortuzar, M., Chin, S. F., Provenzano, E., Bardwell, H. A., Ma, W., Cope, W., Dariush, A., Dawson, S. J., Abraham, J. E., Dunn, J., Hiller, L., Thomas, J., Cameron, D. A., Bartlett, J. M. S., Hayward, L., Pharoah, P. D., Markowetz, F., Rueda, O. M., Earl, H. M., & Caldas, C. (2023). Multi-omic machine learning predictor of breast cancer therapy response. Nature Communications, 14(1), 1234. https://www.nature.com/articles/s41586-021-04278-5
- [8] Cronograma de planificación. CANCER DE MAMA_TFM_2023_rev