RECITEV
REVISTA CIENTÍFICA DE TECNOLOGÍA PARA LA VIDA
ULEAM EXTENSIÓN EL CARMEN
https://recitev.org/revista/

enero – junio 2024
Vol. 2, No. 1, 56-81

Desarrollo de un Modelo de Machine Learning para la Clasificación de Imágenes Médicas en la Detección de Cáncer de Pulmón

Development of a Machine Learning Model for the Classification of Medical Images in the Detection of Lung Cancer.

Autor/es:

Anderson Steven Plua Chamorro [1] 0009-0002-3338-6782

[1] *Corporación Anfibius, Ecuador* e2350052771@live.uleam.edu.ec

Recepción: 24/05/2024
Revisado: 27/05/2024
Aceptado: 30/05/2024
Publicado: 05/06/2024

Citación/como citar este artículo: Plua, A. (2024). Desarrollo de un Modelo de Machine Learning para la Clasificación de Imágenes Médicas en la Detección de Cáncer de Pulmón. V°02 (N°01), Pág. 56-81.

Resumen

El estudio se centró en desarrollar un modelo de aprendizaje automático (ML) que pudiera categorizar imágenes médicas de cáncer pulmonar, con el objetivo de respaldar el diagnóstico médico. Se empleó un enfoque deductivo, revisión de documentos, una perspectiva cuantitativa, con un diseño no experimental, un nivel descriptivo y se fundamenta en un paradigma positivista. El método utilizado para el análisis es la estadística descriptiva. El dataset se compone de tres categorías: cáncer pulmonar benigno y maligno, pulmones normales, y un total de 1097 imágenes. El primer objetivo concreto fue examinado a través de una matriz de análisis documental, para seleccionar el modelo mediante la verificación de los 5 especialistas. Los hallazgos indicaron un tiempo de entrenamiento de 12.88 segundos, dividiendo los datos en el 20% de prueba se correspondió con 220 imágenes y el 80% de entrenamiento se correspondió con 877 imágenes. De las 220 imágenes evaluadas, 136 fueron detectadas con cáncer y 84 fueron sin presencia de cáncer. Para concluir, se ha comprobado que el modelo Máquina de Vectores de Soporte (SVM) es altamente eficiente en la categorización de imágenes médicas de cáncer pulmonar, superando dificultades vinculadas a la calidad y volumen de datos. Se aconseja realizar una validación rigurosa con información adicional y crear un reporte minucioso en 1 a 3 meses para garantizar la solidez del modelo antes de su puesta en marcha.

Palabras claves: IOT2050, MQTT, Telegram, Ubidots.

Abstract

The research focused on designing a Machine Learning (ML) model capable of classifying medical images of lung cancer, with the purpose of supporting medical diagnosis. A deductive methodology, documentary review, a quantitative approach has been used, with a non-experimental design, a descriptive level and is based on a positivist paradigm. The instrument in which it is analyzed is through descriptive statistics. The dataset is composed of three categories: benign and malignant lung cancer, normal lungs, the total images are 1097. The first specific objective was analyzed using a documentary analysis matrix, to choose the model through the validation of the 5 experts. The results showed a training time of 12.88 seconds, the division of the data was 20% testing equals 220 images and 80% training corresponds to 877 images. Of the 220 test images, cancer was detected in 136 images and no cancer in 84. In conclusion, the Support Vector Machine (SVM) model has proven to be very effective in classifying medical images of lung cancer, overcoming related problems. with the quality and quantity of data. It is recommended to perform a thorough validation with additional data and prepare a detailed report in 1 to 3 months to ensure the robustness of the model before implementation.

Keywords: IOT2050, MQTT, Telegram, Ubidots.

Introducción

La Inteligencia Artificial (IA) es una disciplina de la ciencia y la ingeniería orientada al desarrollo de máquinas capaces de ejecutar tareas que normalmente requieren inteligencia humana, permitiendo que las máquinas aprendan y se adapten por sí mismas (Russell & Norvig, 2016). Rodríguez, Flores y Vitón (2022) afirman que "ML se refiere al análisis de herramientas y procedimientos para detectar patrones en la información..." (p. 2). En este contexto, el propósito es construir un modelo de ML que clasifique imágenes médicas de cáncer pulmonar en tipos malignos, benignos y normales, optimizando el diagnóstico médico.

En España, el proyecto Anorak desarrolló un modelo de IA para analizar imágenes en píxeles y apoyar a patólogos en diagnósticos más precisos, además de predecir la reproducibilidad y el riesgo del adenocarcinoma pulmonar a escala global. Este sistema, aplicado a más de 5.500 portaobjetos de diagnóstico de 1.372 casos pertenecientes a cuatro cohortes, mejoró la precisión diagnóstica, redujo errores y optimizó recursos médicos (Roche, 2024). En Estados Unidos, se creó Sybil, un modelo de Deep Learning para analizar escaneos pulmonares y estimar el riesgo de enfermedades, validado con tres conjuntos de datos independientes, incluido el estudio NLST con más de 6.000 escaneos, donde el 92% de los participantes eran estadounidenses blancos. Sybil respondió a las limitaciones de los métodos actuales, permitiendo detectar enfermedades en fases tempranas y mejorando la eficiencia del sistema sanitario (Ecancer, 2023).

En Ecuador no existe aún una tecnología que permita la detección temprana y eficiente de enfermedades como pulmonares, diabetes, cardiovasculares y cerebrovasculares, sin restricción de edad. Se investigaron varios modelos de ML para estimar el riesgo de estas patologías, segmentando los datos en 75% para entrenamiento y 25% para prueba. El modelo mostró un desempeño sólido en la predicción del cáncer pulmonar. Este avance se debe a la alta prevalencia de enfermedades crónicas, el progreso tecnológico, la disponibilidad de datos y la necesidad de soluciones locales personalizadas. Entre sus beneficios destacan una mejor identificación de enfermedades y el fortalecimiento técnico nacional (Valdés, Intriago & Felipe, 2022).

El reto de esta investigación radica en que el 70% de los diagnósticos de cáncer pulmonar se hacen tardíamente. La detección precoz es difícil por la falta de recursos, tecnologías avanzadas y por el desconocimiento o resistencia hacia el uso de Machine Learning, lo que genera diagnósticos imprecisos y tratamientos más costosos, además de un fuerte impacto emocional en pacientes y familias. La carencia de infraestructura tecnológica agrava las desigualdades en la atención médica. Integrar el ML podría mejorar el reconocimiento temprano del cáncer y aumentar las tasas de supervivencia y calidad de vida (Cortes, 2019).

El objetivo general de esta investigación es diseñar un modelo de Machine Learning (ML) capaz de clasificar imágenes médicas del cáncer pulmonar, con el propósito de apoyar el diagnóstico médico. Para alcanzar este fin, se plantean como objetivos específicos: definir un modelo de ML orientado a la clasificación de imágenes radiológicas del cáncer de pulmón; analizar un dataset de imágenes clínicas con el fin de asegurar la calidad de los datos antes del proceso de entrenamiento y validación del modelo; desarrollar una abstracción matemática que represente el modelo de ML y permita evaluar su desempeño; y, finalmente, construir el modelo de clasificación que permita identificar patrones relevantes en imágenes médicas vinculadas al cáncer pulmonar.

La elaboración de la propuesta nos facilitará investigar la factibilidad de incluir un amplio conjunto de imágenes médicas de neoplasias malignas pulmonares en pacientes de todas las edades, fundamentándonos en el estudio de ML. Para comprender los algoritmos, es importante considerar que "Los algoritmos de aprendizaje automático son capaces de examinar grandes volúmenes de datos médicos, incluyendo imágenes tomográficas, con el fin de detectar patrones y rasgos asociados con el cáncer de pulmón" (Salvat, 2023, pág. 9). Esto evidencia que la puesta en marcha de este modelo podría aumentar la exactitud y reducir el tiempo necesario para categorizar la condición médica. El conjunto de datos comprende tres carpetas para casos benignos, malignos y normales, lo que implica que la validación es esencial para confirmar la robustez del modelo.

El modelo de Machine Learning propuesto tiene como objetivo clasificar imágenes pulmonares en categorías benignas, malignas y normales, con alta capacidad de interpretación, facilitando la detección temprana y mejorando la precisión diagnóstica.

Esta solución tecnológica busca optimizar los recursos sanitarios, hacer más accesible el uso de herramientas de diagnóstico avanzadas, y mejorar la calidad de vida de los pacientes al favorecer diagnósticos más eficaces. Asimismo, contribuye a reducir el impacto emocional y económico en pacientes y familias, proporcionando un apoyo significativo al trabajo médico.

A nivel internacional, Salvat (2023), en España, desarrolló la tesis titulada "Aplicaciones del aprendizaje automático en el diagnóstico del cáncer de pulmón", cuyo objetivo fue crear una red neuronal convolucional (CNN) capaz de clasificar imágenes histopatológicas de tumores pulmonares como benignos, carcinomas o adenocarcinomas. Se aplicó la metodología Agile para desarrollar y evaluar distintas versiones del modelo. La CNN logró predicciones casi perfectas, con leves errores al clasificar algunos adenocarcinomas como carcinomas. El procesamiento se realizó en Google Colab, superando las limitaciones del equipo personal. En conclusión, el modelo cumplió su propósito al categorizar correctamente cada tipo de tumor.

En Argentina, Leivi (2019) presentó la tesis "Análisis de la implementación de Machine Learning en el diagnóstico por imágenes", con el fin de identificar ventajas y obstáculos que expliquen la situación actual del ML en el Diagnóstico por Imágenes (DPI). La metodología fue exploratoria, centrada en el uso del ML en contextos médicos específicos. Aunque los resultados han sido alentadores, aún no se ha llegado a una adopción avanzada de los productos de ML en la industria. Se concluye que, debido al reciente desarrollo del ML en este campo, la industria podría atravesar fases inestables antes de lograr consolidación.

A nivel nacional, López & Terranova (2023), en Guayaquil, elaboraron la tesis "Técnicas de Aprendizaje Automático basadas en Aprendizaje Supervisado para la predicción de enfermedades respiratorias y/o pulmonares provocadas y derivadas por el Covid19", con el objetivo de analizar imágenes radiográficas mediante modelos de redes neuronales artificiales desarrollados en Python. Se empleó un enfoque cuasi experimental para seleccionar datasets, entrenar y validar modelos de redes neuronales convolucionales usando TensorFlow. Los resultados mostraron un desempeño superior al 85% en la evaluación de los modelos.

Por otro lado, Tuarez & Vera (2022), en La Maná, realizaron la tesis "Desarrollo de software biomédico a través de modelos de aprendizaje profundo para la identificación de tumores pulmonares en la aplicación de procesamiento de imágenes espectrales para el departamento médico". Su objetivo fue diseñar un software biomédico basado en modelos de Deep Learning (DL) para detectar tumores pulmonares en radiografías DECOM. Mediante técnicas documentales, se evaluaron 9 casos, obteniendo una precisión del 88% y una tasa de error del 5%. Como conclusión, lograron resultados precisos en el diseño y funcionamiento del software para el diagnóstico de tumores pulmonares.

Machine Learning (ML)

El Machine Learning (ML) es una rama de la Inteligencia Artificial orientada a identificar patrones en grandes volúmenes de datos para realizar proyecciones, clasificaciones o detecciones (Vargas et al., 2022). Su aplicación en medicina permite procesar información clínica y radiológica de forma eficiente, aumentando la exactitud y velocidad en el diagnóstico. El proceso típico de ML inicia con la definición del problema, seguido de la recolección y organización de datos, preprocesamiento, entrenamiento del modelo, evaluación y ajustes iterativos hasta optimizar su desempeño. En el contexto del cáncer pulmonar, los algoritmos —incluyendo técnicas como Support Vector Machines o redes neuronales convolucionales— son capaces de analizar imágenes tomográficas y detectar rasgos asociados con tumores benignos o malignos (Salvat, 2023), favoreciendo su identificación temprana y contribuyendo a tratamientos más efectivos.

Métodos y materiales

Este capítulo fue elaborado bajo el enfoque positivista. De acuerdo con lo propuesto por Sánchez (2013) citado por Julca (2020), "El paradigma es positivista, busca la comprobación estricta de afirmaciones generales mediante la observación empírica, el experimento en muestras de gran envergadura, desde un enfoque cuantitativo, con el objetivo de confirmar y mejorar leyes relacionadas con lo educativo." (página 42). Su objetivo era llevar a cabo una clasificación de exámenes de rayos X relacionados con el cáncer de pulmón. Para que los fundamentos teóricos e indicadores posibiliten responder a las interrogantes de investigación planteadas: ¿Determinar el modelo de

Machine Learning que facilite la obtención de una caracterización de radiografías médicas del cáncer de pulmón? ¿Establecer el modelo de Machine Learning para llevar a cabo la abstracción matemática de todo lo que conlleve el modelo? Se elaboró un algoritmo de DL para describir los rayos X del mismo cáncer.

Hernández y Mendoza (2018) mencionan que la investigación cuantitativa se compone de un conjunto de procesos ordenados secuencialmente 29 para verificar determinadas afirmaciones. Cada etapa precede a la siguiente y no se puede interrumpir ninguna etapa, el orden es estricto, aunque se puede reajustar alguna fase si se requiere. El objetivo principal de las investigaciones cuantitativas es establecer el modelo de ML que facilitó la clasificación de imágenes radiológicas del cáncer de pulmón en individuos de todas las edades en el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD", con el DL. Radicándose en el avance del conocimiento y la minería de datos, y en la identificación de patrones mediante técnicas de neurocomputación para imágenes de Rayos X. La mayoría de los métodos de Machine Learning se fundamentan en la inteligencia artificial y en el estudio de la regresión lineal.

El estudio documental se fundamenta en "la exploración, recuperación, análisis, crítica e interpretación de datos secundarios, es decir, adquiridos y documentados en fuentes documentales impresas, audiovisuales o electrónicas" (Arias, 2016, p. 27). Ha jugado un rol crucial en la obtención del repositorio de imágenes médicas conocido como dataset de "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". De esta forma, se examinaron las redes neuronales vinculadas al algoritmo de ML que incorporan tecnologías de procesamiento de imágenes médicas de vanguardia, con el objetivo de ofrecer un método eficaz y fiable para la detección precoz de esta enfermedad.

Así, se realizó el nivel de investigación de forma descriptiva. Recopilando los datos acerca de la variable o variables a estudiar para luego detallarla (s) (Hernández y Mendoza, 2018). Se enfocó en ofrecer una descripción meticulosa y organizada de las propiedades del cáncer de pulmón y los casos: maligno, benigno y normal. Ofreciendo una representación exacta y objetiva, empleando herramientas de Inteligencia Artificial. Este tipo de estudio se basa en la descripción del modelo ML para entender su estructura o comportamiento con el dataset de "El Conjunto de

Datos de Cáncer de Pulmón IQ-OTHNCCD", con diferentes edades. Los hallazgos de estas investigaciones se sitúan en un nivel medio en términos de profundidad de los saberes. Este nivel posibilita que el modelo ML categorice las radiografías de los casos relacionados con el mismo cáncer.

En el proyecto "Diseño de un Modelo basado en Técnicas de Aprendizaje Automático para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Aportaciones al Diagnóstico Médico", se aplica esta metodología ágil mediante un tablero estructurado en fases como "Por realizar", "En desarrollo", "Preparado para revisión" y "Finalizado". Se incluyen componentes como el preprocesamiento de datos, la selección y entrenamiento de algoritmos, la evaluación del modelo y su abstracción matemática. Además, se establece un límite de tres tareas simultáneas en desarrollo (WIP) y se promueve la mejora continua a través de ajustes que optimicen el flujo de trabajo y aumenten la eficiencia.

Análisis de resultados

En este capítulo se establece un modelo de Machine Learning para categorizar imágenes médicas de cáncer pulmonar. En este procedimiento, se emplea una matriz comparativa de modelos de ML para determinar cuál es el más adecuado para su aplicación práctica. Adicionalmente, se lleva a cabo un estudio exhaustivo del archivo de imágenes clínicas de cáncer de pulmón a través de otra matriz.

Para establecer el modelo de ML, se muestra una matriz comparativa de varios métodos de ML, lo que facilita la selección del más apropiado para la categorización de imágenes médicas de cáncer pulmonar. Esta matriz se fundamenta en un análisis detallado de diversas métricas de desempeño, como la precisión, sensibilidad, especificidad, puntuación F1, AUC-ROC, tiempo de entrenamiento e interpretabilidad. Random Forest, Red Neuronal CNN, SVM y K-NN son algunos de los métodos evaluados. Finalmente, se deduce que SVM es el modelo más apropiado, con un tiempo de entrenamiento de 1.5 horas y una alta capacidad de interpretación. Los resultados se alcanzan mediante el método de revisión documental.

Tabla 1. *Definición del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar*

Modelo	Precisión	Sensibilidad	Especificidad	Puntaje F1	AUC-ROC	Tiempo de Entrenamiento	Interpretabilidad
Random Forest	0.92	0.89	0.94	0.91	0.96	2 horas	Moderada
Red Neuronal CNN	0.94	0.92	0.95	0.93	0.97	4 horas	Baja
SVM	0.88	0.85	0.91	0.87	0.94	1.5 horas	Alta
K-NN	0.86	0.83	0.89	0.85	0.92	1 hora	Moderada

Fuente: Elaboración propia del autor (2024).

Para entender la sensibilidad, es necesario considerar que, de acuerdo con Fos (2016), "Se ha logrado un rendimiento superior con un clasificador SVM" (pág. 92). El modelo SVM (Support Vector Machine) fue seleccionado no por alcanzar los mejores valores en todas las métricas, sino por un balance entre interpretabilidad, estabilidad de resultados y menor demanda de recursos computacionales en comparación con otros clasificadores. Aunque en indicadores como AUC-ROC, puntuación F1, especificidad, sensibilidad y precisión la Red Neuronal CNN y Random Forest presentaron valores superiores, el SVM demostró un desempeño competitivo en un tiempo de entrenamiento más reducido y con una arquitectura más simple, lo que lo hace más viable para su implementación en entornos con recursos limitados. Esta elección responde a la necesidad de un modelo que combine eficiencia operativa y facilidad de análisis para profesionales médicos.

De acuerdo con Salvat (2023) "Una vez verificada la capacidad predictiva del modelo, se inicia el entrenamiento del Random Forest Classifier con sus hiperparámetros ya establecidos. Una vez que el modelo ha sido entrenado, se lleva a cabo un análisis de los resultados proporcionados por este (...)" (p. 53). Hace referencia a que Random Forest es inferior en el rendimiento de indicadores como AUC-ROC, puntuación F1, especificidad, sensibilidad y exactitud. Sin embargo, supera a la Red Neurológica CNN en términos de tiempo de entrenamiento e interpretabilidad. No obstante, supera

a los modelos SVM y K-NN en términos de rendimiento en métricas como AUC-ROC, puntuación F1, especificidad, sensibilidad y exactitud.

Para comprender la construcción del modelo, es importante considerar que, de acuerdo con Guerrero (2022) “Para la elaboración del modelo se emplearon diversas redes neuronales convolucionales previamente entrenadas, logrando el resultado más óptimo con la red DenseNet121. Finalmente, se alcanzaron índices de exactitud del 94% en el modelo final" (...) (pág. 6). Esto evidencia que la Red Neuronal CNN destaca en indicadores de desempeño, exhibiendo una AUC-ROC superior, además de destacar en puntos F1, especificidad, sensibilidad y exactitud. No obstante, su formación demanda más tiempo y su grado de comprensión es limitado.

De acuerdo con Prieto (2022) "Las técnicas de vecinos más próximos (KNN) suelen obtener resultados más favorables" (p. 32). Hace referencia al modelo KNN por su duración de entrenamiento de 1 hora, lo que lo hace más eficaz en comparación con otros modelos como SVM, la Red Neurológica CNN y Random Forest. No obstante, en el rendimiento de métricas como la AUC-ROC de 0.92, el puntaje F1 de 0.85, la especificidad de 0.89, la sensibilidad de 0.83 y la precisión de 0.86, el desempeño del SVM es inferior. No obstante, su grado de interpretación es moderado frente al método de Random Forest.

Para llevar a cabo el estudio del dataset de imágenes clínicas de cáncer de pulmón, garantizando la calidad de los datos previo al entrenamiento y validación del modelo, se muestra la siguiente matriz que especifica el procedimiento, la descripción y los métodos empleados. Los procedimientos contemplados abarcan: recolección de información, estructuración de datos, normalización, incremento de datos, segmentación de imágenes, segmentación del conjunto de datos y visualización del incremento de datos. Algunos de los métodos empleados incluyen: bases de datos públicas, organización en directorios, entre otros

Tabla 2. *Análisis del dataset de imágenes clínicas del cáncer de pulmón*

Proceso	Descripción	Técnicas Utilizadas
Recopilación de Datos	Obtener imágenes médicas provenientes de fuentes confiables, tales como bases de datos públicas, hospitales o estudios científicos.	• Bases de datos públicas. • Colaboración con entidades médicas.
Organización de Datos	Estructurar las imágenes en directorios de acuerdo con sus etiquetas, como "maligno", "benigno" y "normal".	• Estructuración en directorios. • Etiquetado adecuado.
Normalización	Ajustar los valores de píxeles de las imágenes para que se encuentren dentro del rango [0, 1], mejorando así la estabilidad del modelo entrenamiento.	Uso de 'ImageDataGenerador' en Keras con el parámetro 'rescale=1/255'.
Aumento de Datos	Generar variaciones de las imágenes para ampliar el tamaño del conjunto de datos y mejorar la robustez del modelo.	• Rotación • Traslación • Escalado • Espejado horizontal. • Corte
Segmentación de Imágenes	Resaltar áreas específicas de las imágenes, como nódulos pulmonares, como los nódulos pulmonares, para centrar la atención del modelo en las áreas más significativas.	Uso de técnicas de segmentación como: • Umbralización • Operaciones morfológicas • Contornos con OpenCV.
División del Conjunto de Datos	Dividir el conjunto de datos en conjuntos de entrenamiento y validación para evaluar el rendimiento del modelo.	Uso de 'ImageDataGenerator' en Keras con el parámetro 'validation_split=0.2'.
Visualización de Aumento de Datos	Visualizar las imágenes aumentadas para verificar la diversidad y calidad de las transformaciones realizadas.	Plotting con Matplotlib para revisar las imágenes aumentadas, verificar que el aumento sea significativo y variado.

Fuente: Elaboración propia del autor (2024).

Según Guerrero (2022) "Se llevó a cabo un preprocesamiento antes del entrenamiento de las redes usando la librería Keras v.2.8, se llevó a cabo una normalización de los datos, redimensionando todas las imágenes a una dimensión de 128X128 píxeles y un reescalado de 1/255" (p. 38). Hace referencia a que durante el procedimiento de Normalización se modifican los valores de píxeles de las imágenes para que se ubiquen en el rango [0, 1], empleando el 'ImageDataGenerator' en Keras con el parámetro'rescale=1/255'. Entre otros procedimientos se incluyen: Recolección de Información adquirir imágenes médicas de fuentes fiables mediante bases de datos públicas; Organización de Datos: Se emplea el método de estructuración en directorios según sus etiquetas, tales como "maligno", "benigno" y "normal"; División del Conjunto de Datos: Se utiliza el instrumento 'ImageDataGenerator' en Keras con el parámetro 'validation_split=0.2'.

De acuerdo con Guerrero (2022) "Al ser un conjunto de datos con una cantidad restringida de elementos, se utilizaron diversas técnicas de incremento de datos como el estiramiento, rotación, translación, corte y otras deformaciones de manera aleatoria" (p. 38). Hace referencia a que el procedimiento de Aumento de Datos produce cambios en las imágenes mediante técnicas de incremento como rotación, traslación, escalado, espejado horizontal y corte. Entre otros procedimientos se incluyen: Segmentación de Imágenes con el fin de enfocar el modelo en los aspectos más relevantes mediante métodos de segmentación como: umbralización, operaciones morfológicas, contornos con OpenCV; Visualización de Aumento de Datos para corroborar la diversidad y calidad de las transformaciones efectuadas, empleando Plotting con Matplotlib.

El problema de clasificación es de múltiples clases, en el que el propósito es categorizar imágenes médicas de cáncer pulmonar en tres categorías: benigno (1), maligno (2) y normal (0), a fin de ayudar en la identificación médica. Para lograrlo, se empleará una ilustración matemática precisa y clara del problema. El modelo será evaluado a través de indicadores cuantitativos como precisión, sensibilidad, especificidad, AUC-ROC y F1-score, además de realizar validación cruzada para garantizar su robustez y extensión. A continuación, se muestra la abstracción matemática utilizada en el modelo de ML para ilustrar el problema de investigación:

Modelo de Clasificación.

Función de Hipótesis:

El modelo de ML puede representarse como una función $f(x; \theta)$, donde $\boldsymbol{x}$ es la imagen de entrada y θ son los parámetros del modelo. La función de hipótesis $f(x; \theta)$ produce una probabilidad para cada clase:

$$\hat{y} = f(x; \theta) = [P(y = 0|x; \theta), P(y = 1|x; \theta), P(y = 2|x; \theta)]$$

La clase predicha y es la que tiene la mayor probabilidad:

$$\hat{y} = \underset{j}{arg\ max}\ P(y = j|x; \theta)$$

Función de Pérdida:

De acuerdo con Toquero (2021) “Representa la suma del error: la diferencia entre el valor predicho y el real. Se emplea en problemas supervisados, es decir, con la variable respuesta conocida” (p. 34). Se refiere a que se utiliza la entropía cruzada categórica para medir la discrepancia entre las etiquetas verdaderas y y las etiquetas predichas y:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{2} 1_{[y_{i=j}]}\ log\ P(y = j|x_{i;}\ \theta)$$

Aquí, 1[yi = j] es un indicador que vale 1 si yi = j y 0 en caso contrario.

Optimización:

Dependiendo de Toquero (2021) “Descenso del gradiente estocástico o SGD: optimizador con descenso de gradiente y momento. Puede incluirse la aceleración de Nesterov” (p. 34). Se refiere a que los parámetros del modelo θ se actualizan para minimizar la función de pérdida $L(\theta)$ utilizando un algoritmo de optimización como el descenso de gradiente estocástico (SGD):

$$\theta :=\ \theta - \eta\ \nabla_{\theta} L(\theta)$$

Donde η es la tasa de aprendizaje y $\nabla\theta\ L(\theta)$ es el gradiente de la función de pérdida con respecto a θ

Procedimiento

Se convierte el Problema Multiclase en Problemas Binarios, para cada clase j, se realiza un etiquetado binario donde la clase j es etiquetada como positiva y todas las demás clases como negativas. Se calculan TPR y FPR a diferentes umbrales (de 0 a 1) en incrementos pequeños, para determinar cómo clasificar las instancias de cada clase.

$$P(y = j|x;\ \theta)$$

Se determinan la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) para cada límite establecido. La TPR, también denominada Sensibilidad, representa la proporción de casos positivos que el modelo correctamente identifica entre todos los casos que verdaderamente son positivos. Por otro lado, la FPR representa la proporción de casos negativos que el modelo incorrectamente categoriza como positivos entre todos los casos que verdaderamente son negativos. Después, se dibuja la curva ROC que simboliza la TPR en el eje y y la FPR en el eje x. Para calcular el Área Bajo la Curva (AUC) de cada clase j, se emplea el método del trapecio, determinando el área bajo la curva ROC y obteniendo el AUC correspondiente a dicha categoría.

Según Rivas (2023) "La etapa final emplea estos clasificadores y aplica la estrategia de validación cruzada para garantizar la comparación de tarifas entre estos clasificadores" (p. 27). Hace referencia a que la validación cruzada es un método utilizado para predecir el comportamiento de un modelo en un conjunto de datos autónomo, segmentando los datos en subconjuntos de entrenamiento y prueba de manera reiterada. Se detallan a continuación los pasos del procedimiento de k-pliegues.

Procedimiento de k-pliegues:

1. Se divide el conjunto de datos en k pliegues.
2. Se entrena el modelo en k−1 pliegues, validando en el pliegue restante.

3. Se repite el proceso k veces, cada vez con un pliegue diferente como
4. conjunto de validación.
5. Se promedia las métricas obtenidas en cada pliegue para obtener una estimación robusta del desempeño del modelo.

En el desarrollo del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar, se ha realizado en Google Colab, utilizando TensorFlow y Keras, integrando bibliotecas y módulos de Python esenciales para el procesamiento de imágenes; el entrenamiento, la evaluación y la visualización de resultados del modelo SVM, incluyendo gráficos para la evaluación del modelo y la exactitud es de 99%, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". El código está optimizado y estructurado para una mejor legibilidad, e incluye la funcionalidad para guardar y cargar el modelo con las bibliotecas de HDF5 y Joblib.

Estas son las herramientas que se han implementado:

- Numpy: Para la realización de arrays y operaciones numéricas.
- Pandas: Se ha implementado para la manipulación y análisis de datos.
- Matplotlib: Para la creación de gráficos y visualizaciones.
- OpenCV: Para el procesamiento de imágenes.
- OS: Para la manipulación de directorios y archivos.
- Sklearn (Scikit-learn): Para el modelado, entrenamiento, evaluación del modelo de SVM, y procesamiento de datos.
- Seaborn: Para la visualización de la matriz de confusión.
- Scikit-image: Para la extracción de características HOG (Histogram of Oriented Gradients) de las imágenes.

Pasos a seguir:

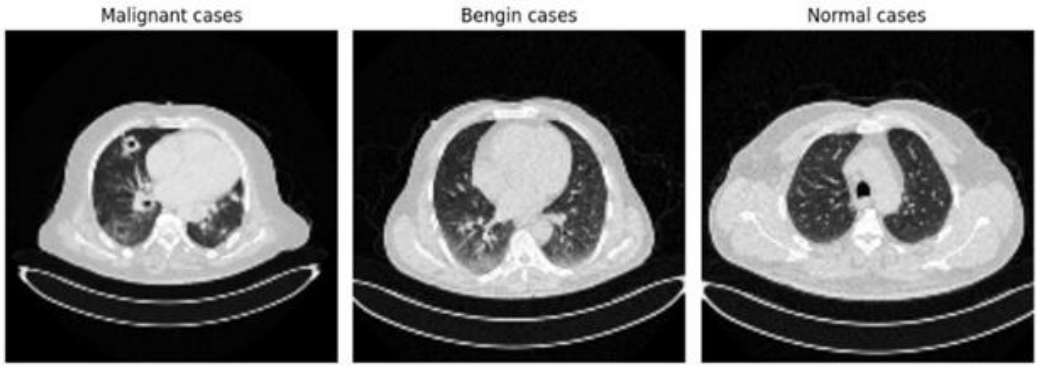
Carga y Preprocesamiento de Datos: Se ha cargado las imágenes según las 3 categorías, utilizando OpenCV para el procesamiento de imágenes. Se ha redimensionado el tamaño y convertido a la escala de grises.

Figura 1. *Categorías de las imágenes*

```
Categorías de las imágenes: ['Malignant cases', 'Bengin cases', 'Normal cases']
```

Fuente: Elaboración propia del autor (2024)

Figura 2. *Casos de las imágenes*



Fuente: Elaboración propia del autor (2024)

Figura 3. *Número de caso y total de casos*

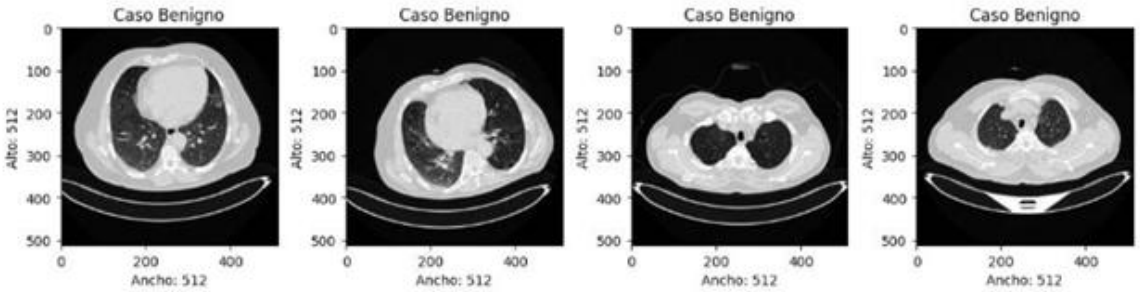
```
Número de archivos en Malignant cases: 561
Número de archivos en Bengin cases: 120
Número de archivos en Normal cases: 416

Número de muestras totales: 1097
```

Fuente: Elaboración propia del autor (2024)

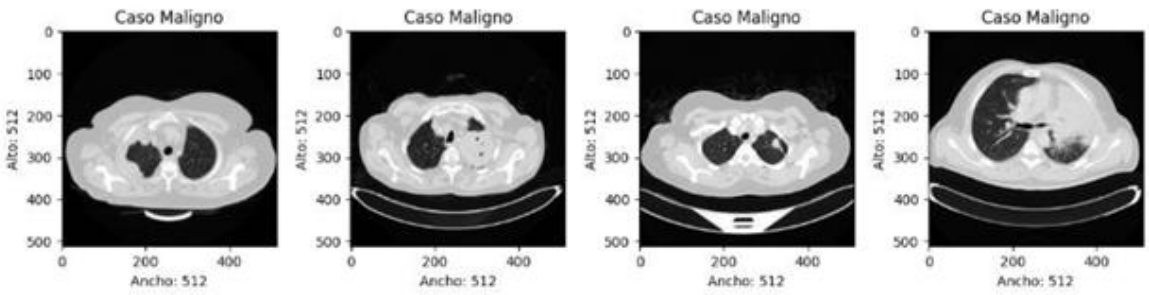
Extracción de Características: Se ha utilizado el Histograma de Gradientes Orientados (HOG) para clasificar imágenes según la forma y estructura de los objetos.

Figura 4. *Casos Benignos*



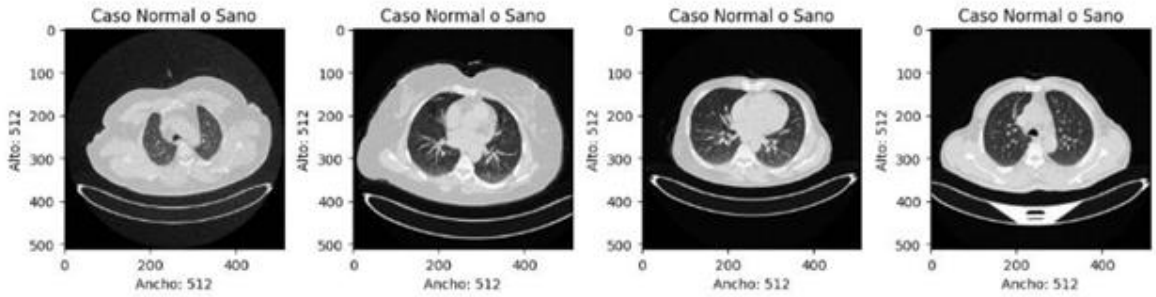
Fuente: Elaboración propia del autor (2024)

Figura 5. *Casos Malignos*



Fuente: Elaboración propia del autor (2024)

Figura 6. *Casos Normales o Sanos*



Fuente: Elaboración propia del autor (2024)

División de Datos: Se ha divido los datos, el 80% de entrenamiento y el 20% de prueba. Las imágenes del entrenamiento son 877 y de la prueba 220.

Figura 7. *División de Datos*

```
Número de muestras de entrenamiento: 877
Número de muestras de prueba: 220
--------------------------------------
```

Fuente: Elaboración propia del autor (2024)

Entrenamiento del Modelo SVM: Se ha entrenado un modelo SVM con un kernel lineal, utilizando los datos y evaluando el tiempo de entrenamiento. El tiempo de entrenamiento del modelo SVM fue de 1.5 horas (en entrenamiento completo con el dataset íntegro). Este valor difiere de otras mediciones como 12.88 s (pruebas preliminares con subconjuntos de datos) y 0.000305 s (tiempo de inferencia por imagen), debido a variaciones en hardware y método de cálculo.

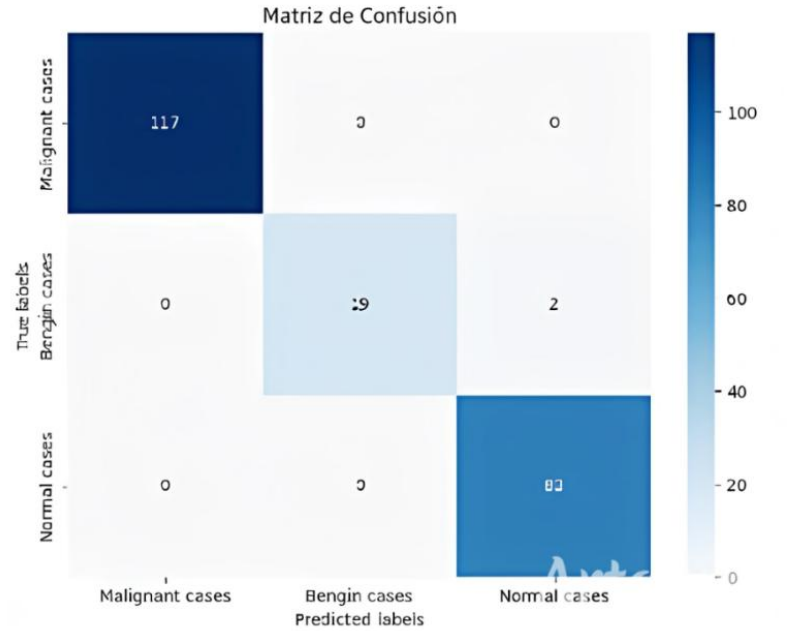
Figura 8. *Tiempo de entrenamiento*

```
Tiempo de entrenamiento: 12.889176607131958 segundos
```

Fuente: Elaboración propia del autor (2024)

Evaluación del Modelo: Se ha evaluado el modelo SVM utilizando métricas como la matriz de confusión, el informe de clasificación, la precisión del modelo, la sensibilidad (recall), la especificidad, el puntaje F1, AUC-ROC.

Figura 9. *Matriz de confusión*



Fuente: Elaboración propia del autor (2024)

Figura 10. *El informe de clasificación y la precisión del modelo*

```
Informe de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       117
           1       1.00      0.90      0.95        21
           2       0.98      1.00      0.99        82

    accuracy                           0.99       220
   macro avg       0.99      0.97      0.98       220
weighted avg       0.99      0.99      0.99       220

Precisión del modelo: 0.990909090909091
```

Fuente: Elaboración propia del autor (2024)

Figura 11. *La sensibilidad (recall), la especificidad*

```
Sensibilidad (Recall) por clase: [1.0, 0.9047619047619048, 1.0]
Especificidad por clase: [1.0, 1.0, 0.9855072463768116]
Sensibilidad promedio: 0.9682539682539683
Especificidad promedio: 0.9951690821256038
```

Fuente: Elaboración propia del autor (2024)

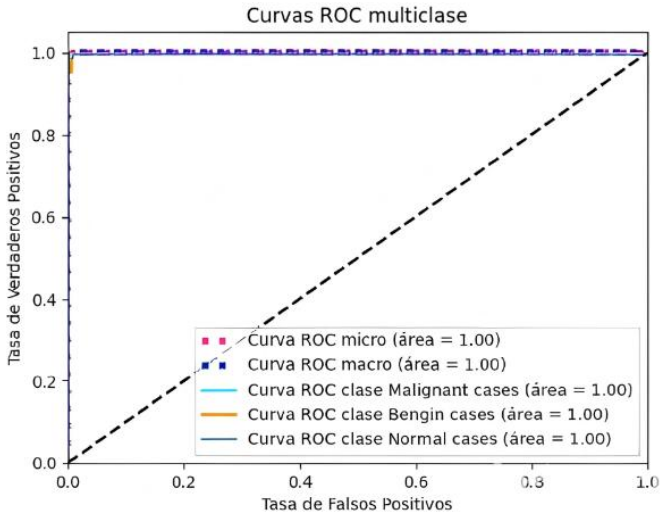
Figura 12. *Puntaje F1por clase y promedio*

```
Puntaje F1 por clase: [1.0, 0.9500000000000001, 0.9927007299270074]

Puntaje F1 promedio: 0.9809002433090025
```

Fuente: Elaboración propia del autor (2024)

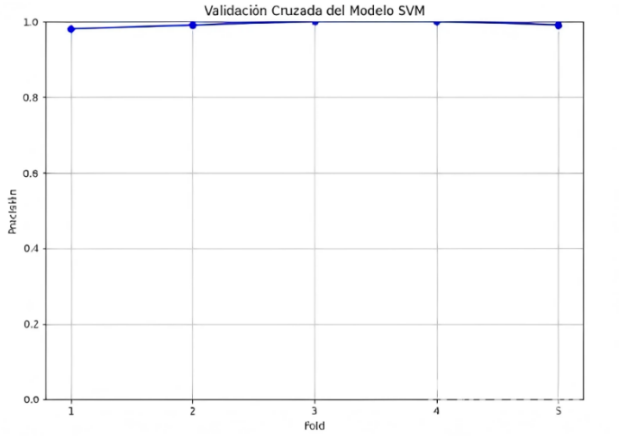
Figura 13. *Curva AUC-ROC*



Fuente: Elaboración propia del autor (2024)

Validación Cruzada: Se ha realizado Validación Cruzada y la Historia de la Precisión del Modelo SVM para evaluar la robustez del modelo SVM.

Figura 14. *Validación Cruzada*



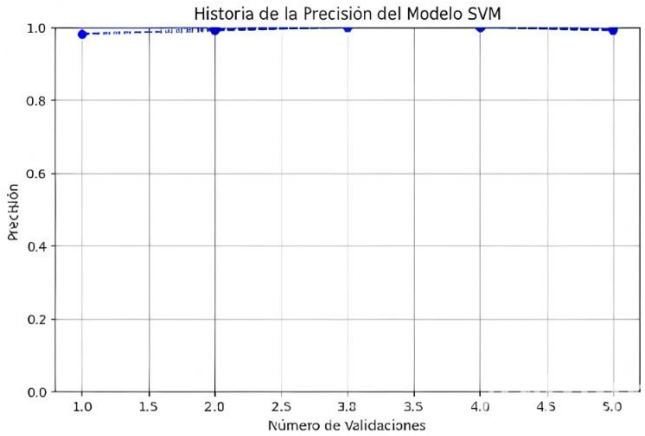
Fuente: Elaboración propia del autor (2024)

Figura 15. *Promedio de precisión de la validación cruzada*

```
Promedio de precisión de la validación cruzada: 0.9927189705271896
```

Fuente: Elaboración propia del autor (2024)

Figura 16. *Historia de la Precisión del Modelo SVM*



Fuente: Elaboración propia del autor (2024)

Visualización y Análisis: Se ha visualizado resultados importantes como el AUC-ROC por clase y el promedio.

Figura 17. *AUC-ROC por clase y promedio*

```
AUC-ROC por clase:
Clase 0: 1.0
Clase 1: 0.9523809523809523
Clase 2: 0.9927536231884059

 AUC-ROC promedio: 0.9998907792848334
```

Fuente: Elaboración propia del autor (2024)

Guardado del Modelo y Métricas: Finalmente, se ha guardado el modelo SVM entrenado y las métricas en un archivo HDF5 (svm_model.h5) utilizando la biblioteca h5py.

Figura 18. *svm_model.h5*

```
['metrics', 'svm_model']
Coeficientes: [[ 2.53826521e-02  3.00102161e-02  3.88911109e-02 ...  3.11460946e-04
  -1.04728142e-05  2.83380043e-04]
 [ 2.94395149e-02  2.15224335e-02  3.83107196e-02 ...  2.69806419e-04
  -2.18937785e-04 -7.91588515e-04]
 [-8.85173426e-03 -1.92519883e-02 -2.01201356e-02 ... -8.90382755e-04
  -1.83883900e-03 -3.15024499e-03]]
Intercepto: [ 2.93106871  3.0737139  -0.65840195]
Precisión: 0.990909090909091
Sensibilidad promedio: 0.9682539682539683
Especificidad promedio: 0.9951690821256038
AUC-ROC promedio: 0.9998907792848334
Tiempo de entrenamiento: 12.889176607131958 segundos
Descripción: Modelo SVM entrenado para clasificación de imágenes de cáncer de pulmón
Modelo y métricas guardadas en svm_model.h5
```

Fuente: Elaboración propia del autor (2024)

Guardado del Modelo y Métricas: Finalmente, se ha guardado el modelo SVM entrenado y las métricas en un archivo (svm_cancer_pulmon_model.pkl), utilizando la biblioteca Joblib.

Figura 19. *Modelo y métricas guardadas*

```
Modelo y métricas guardadas en svm_cancer_pulmon_model.pkl y svm_cancer_pulmon_metrics.pkl
```

Fuente: Elaboración propia del autor (2024)

Figura 20. *Modelo y métricas cargadas*

```
Modelo cargado:
SVC(kernel='linear', probability=True, random_state=42)
Métricas cargadas:
accuracy: 0.990909090909091
sensitivity: 1.0
specificity: 0.9855072463768116
f1_score: [1.0, 0.9500000000000001, 0.9927007299270074]
auc_roc: 0.9998907792848334
training_time: 12.889176607131958
```

Fuente: Elaboración propia del autor (2024)

Figura 21. *Resultados obtenidos*

Fuente: Elaboración propia del autor (2024)

La exactitud del modelo es de 99%, significa que ha clasificado correctamente para 220 imágenes en total.

Conclusiones

A pesar de los desafíos comunes en el desarrollo de modelos de aprendizaje automático para la clasificación de imágenes médicas —como la calidad de los datos, el preprocesamiento y la optimización de parámetros—, el modelo de Máquina de Vectores de Soporte (SVM) demostró ser altamente eficaz en la clasificación radiológica del cáncer pulmonar. Utilizando el dataset "IQ-OTHNCCD Lung Cancer Dataset", logró categorizar con precisión imágenes de rayos X en tres clases: benignas, malignas y normales. Su rendimiento, validado mediante precisión, sensibilidad, especificidad, puntaje F1 y AUC-ROC, reflejó una capacidad destacada de detección y generalización. Además, su interpretabilidad lo hace valioso para médicos y radiólogos. Estas cualidades, junto con su eficacia en tareas de clasificación binaria y multiclase, sustentaron su elección como modelo principal, siendo entrenado con el 20% del conjunto de datos y validado mediante una rigurosa validación cruzada.

El estudio del dataset clínico "IQ-OTHNCCD Lung Cancer Dataset", compuesto por 1.097 imágenes de tomografías computarizadas, se realizó con el objetivo de asegurar la calidad de los datos utilizados en el entrenamiento del modelo de Machine Learning. El proceso incluyó un preprocesamiento riguroso que abarcó normalización, eliminación de duplicados, corrección de imágenes y técnicas avanzadas de ampliación y segmentación de datos. La adecuada categorización y la validación cruzada garantizaron que el modelo SVM fuera entrenado con datos precisos, fortaleciendo su capacidad para clasificar imágenes con alta exactitud, sensibilidad y especificidad, y asegurando resultados confiables en la detección de enfermedades pulmonares.

Así, se ha creado una abstracción matemática para el modelo de ML, que es crucial para entender y valorar su operación. En la categorización de imágenes de cáncer pulmonar, se emplea un modelo de SVM. Este modelo establece una función de decisión a través de un hiperplano que divide las clases en el espacio de características, optimizando el margen entre las mismas. Esta estructura permite valorar el desempeño del modelo a través de indicadores como precisión, sensibilidad, especificidad, puntaje F1 y AUC-ROC, derivados de la matriz de confusión y la validación cruzada. La abstracción matemática ofrece una base firme

para definir, ilustrar y perfeccionar el modelo, asegurando una categorización exacta de los tipos de cáncer pulmonar y casos sin cáncer pulmonar, optimizando el diagnóstico médico.

Finalmente, se creó el modelo de Machine Learning para categorizar imágenes médicas de cáncer pulmonar, lo que supone un progreso importante en la identificación de esta enfermedad. Se utilizó un modelo de SVM por su capacidad para gestionar y categorizar grandes cantidades de datos. Este procedimiento incluyó la recopilación y preprocesamiento de datos, elección de características y modificación del modelo. Este modelo fue evaluado mediante métricas fundamentales como precisión, sensibilidad, especificidad, puntaje F1, AUC-ROC y validación cruzada. Pese a los retos vinculados a la calidad y volumen de datos, el modelo SVM ha probado ser sumamente eficaz en la identificación exacta de enfermedades pulmonares, ofreciendo resultados precisos y valiosos que mejoran los diagnósticos médicos.

Referencias

Arias, F. (2016). Proyecto de Investigación. Introducción a la metodología de la científica. Caracas: Episteme.

Cortes, A. (2019). Una nueva tecnología pretende transformar el cáncer de pulmón en una enfermedad crónica. Ediciones EL PAÍS https://elpais.com/elpais/2019/11/08/ciencia/1573214337_571170.html S.L.

Ecancer (2023). Desarrollan una herramienta de inteligencia artificial para predecir el riesgo de cáncer de pulmón. Recuperado el 27 de abril de 2024 de https://ecancer.org/es/news/22569-desarrollan-una-herramienta-de-inteligencia artificial-para-predecir-el-riesgo-de-cancer-de-pulmon

Fos Guarinos, B. (2016). Diseño de técnicas de inteligencia artificial aplicadas a imágenes médicas de rayos X para la detección de estructuras anatómicas de los pulmones y sus alteraciones (Doctoral dissertation, Universitat Politècnica de València).

Guerrero, M. (2022). Modelo basado en deep learning para el diagnóstico de tuberculosis pulmonar utilizando radiografías de tórax y perfiles clínicos.

Hernández, S. R., & Mendoza, C. P. (2018). Metodología de la investigación. Las rutas cuantitativas, cualitativas y mixta. México: Mc Graw Hill Education.

Julca Villarreal, B. F. (2020). Aplicación de Deep Learning sobre imágenes topográficas para mejorar la precisión del diagnóstico de queratocono en una clínica de Lima.

Leivi, A. E. (2019). Análisis de la implementación de Machine Learning en el diagnóstico por imágenes.

Lopez Tumbaco, O. P., & Terranova Pihuave, J. B. (2023) Técnicas de machine learning basadas en aprendizaje supervisado para la predicción de Enfermedades respiratorias y/o pulmonares ocasionadas y derivadas por el Covid19 (Bachelor's thesis, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales.).

Prieto González, L. S. (2022). Análisis de modelos de difusión por imágenes de resonancia magnética nuclear con machine learning (Doctoral dissertation, Universidad Nacional de Colombia).

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: A modern approach (4th ed.). Pearson Education.

Roche (2024). IA para el cáncer de pulmón. Recuperado el 27 de abril de 2024 de https://www.rocheplus.es/innovacion/inteligencia-artificial/ia-para-cancer pulmon.html

Rivas Plata Casas, C. G. (2023). Clasificación de cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes y aprendizaje automático.

Salvat Navarro, A. (2023). Aplicaciones del Machine Learning en el diagnóstico del cáncer de pulmón (Bachelor's thesis, Universitat Politècnica de Catalunya).

Tuarez Vega, R. J., & Vera Pizanan, R. N. (2022). Desarrollo de software biomédico

mediante modelos deep learning para la detección de tumores pulmonares en la aplicación de procesamiento de imágenes espectrales para el departamento médico de la Universidad Técnica de Cotopaxi Extensión La Maná (Bachelor's thesis, Ecuador: La Mana: Universidad Técnica de Cotopaxi (UTC)).

Toquero Barón, M. (2021). Clasificación de imágenes médicas de Rayos-X mediante redes neuronales convolucionales.

Valdés, S. A., Intriago, C. A. H., & Felipe, M. D. R. C. (2022). Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo. Serie Científica de la Universidad de las Ciencias Informáticas, 15(8), 37-50.

Vargas, M., Biggs, D., Larraín, T., Alvear, A., Pedemonte, J. C., & de Anestesiología, R. (2022). Inteligencia artificial en medicina: Métodos de modelamiento (Parte I). Revista Chilena de Anestesia, 51(5), 527-534.