FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

Clasificación de cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes y aprendizaje automático

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Autor (es)
Bach. Rivas Plata Casas Carlos Gualberto

ORCID https://orcid.org/0000-0002-9249-5586

Asesor(a)
Dr. Tuesta Monteza Victor Alexci

ORCID https://orcid.org/0000-0002-5913-990X

Línea de Investigación
Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú
Año 2023

CLASIFICACIÓN DE CÁNCER DE PULMÓN EN IMÁGENES DE TOMOGRAFÍAS MEDIANTE PROCESAMIENTO DE IMÁGENES Y APRENDIZAJE AUTOMÁTICO

APROBADO

Aprobación del jurado

DR. CHIRINOS MUNDACA CARLOS ALBERTO

Presidente del Jurado de Tesis

MG. CACHAY MACO JUNIOR EUGENIO

Secretario del Jurado de Tesis

MG. MINGUILLO RUBIO CÉSAR AUGUSTO

Vocal del Jurado de Tesis

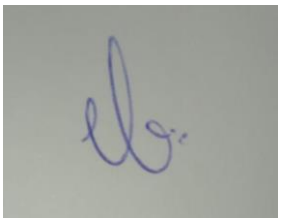
DECLARACIÓN JURADA DE ORIGINALIDAD

Quien(es) suscribe(imos) la **DECLARACIÓN JURADA**, soy(somos) Carlos Gualberto Rivas Plata Casas del Programa de Estudios de Ingeniería de sistemas de la Universidad Señor de Sipán S.A.C, declaro (amos) bajo juramento que soy (somos) autor(es) del trabajo titulado:

Clasificación de cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes y aprendizaje automático

El texto de mi trabajo de investigación responde y respeta lo indicado en el Código de Ética del Comité Institucional de Ética en Investigación de la Universidad Señor de Sipán (CIEI USS) conforme a los principios y lineamientos detallados en dicho documento, en relación a las citas y referencias bibliográficas, respetando al derecho de propiedad intelectual, por lo cual informo que la investigación cumple con ser inédito, original y autentico.

En virtud de lo antes mencionado, firman:

Rivas Plata Casas Carlos Gualberto	DNI: 72932482	

Pimentel, 25 de 04 de 2023.

Dedicatoria

Dedicado a mis familiares, amigos, y profesores, esta tesis es una prueba de mi gratitud por el amor, el apoyo y el aliento que me han dado durante el largo camino hacia la realización de este logro. Nunca habría podido conseguir esto sin sus esfuerzos incansables. Sus palabras de ánimo y su capacidad de creer en mí incluso cuando yo no lo hacía, me han ayudado a superar los desafíos y lograr mis objetivos. De corazón, agradezco cada momento de apoyo que me han brindado.

Agradecimientos

Quiero agradecer a todas las personas que han ayudado a hacer este proyecto una realidad. En primer lugar, quiero agradecer a mi familia por su apoyo incondicional y por creer en mí a pesar de los obstáculos en el camino. También quiero agradecer a mis amigos y compañeros de clase por escucharme, comprenderme y apoyarme. A mis profesores, por sus conocimientos y exigencia que formaron un profesional competente y por último agradecer a todos aquellos que me han ayudado a superar las dificultades que he enfrentado para llegar a este punto.

Índice

Índice de tablas

Índice de figuras

Resumen

La detección de cáncer de pulmón puede resultar complicada para los profesionales de la salud en sus primeras etapas, ya que es difícil identificarlo a partir de imágenes médicas, lo que supone un obstáculo para comenzar un tratamiento adecuado para los pacientes. Esta enfermedad es la principal causa de muerte, con un incremento de nuevos casos, fallecimientos y cada año mueren más personas por este cáncer que por cáncer de mama, próstata y colon. Las técnicas de clasificación tradicionales tienden a no mejorar sus métricas de evaluación debido a sus procesos de filtrado, segmentación, extracción de características y clasificación. La detección tradicional requiere una gran cantidad de tiempo y recursos económicos. La metodología consta de seis pasos: se inicia con una investigación previa para revisar diferentes estudios. Luego, se selecciona un conjunto de datos. En la tercera etapa se eligen las arquitecturas más destacadas para clasificar con relación al conjunto de datos ImageNet. La cuarta etapa se configuran los modelos para entrenamiento y validación. La quinta etapa se evalúa el consumo de recursos y rendimiento de los modelos. Finalmente, se crea una aplicación web que emplea la arquitectura con los mejores resultados. Después de analizar las arquitecturas seleccionadas se obtuvo métricas porcentuales de 97% o más. Sin embargo, las pruebas revelaron que las métricas de exactitud y precisión alcanzaron porcentajes de 95% y 91%, respectivamente. En conclusión, Efficientb4_DA logra mejores resultados alcanzando una exactitud de 95.32%, una precisión de 91.29%, una sensibilidad de 89.84% y una puntuación F de 90.54%.

Palabras Clave: Cáncer de pulmón, detección, clasificación, aprendizaje automático, métricas de evaluación.

Abstract

The detection of lung cancer can be complicated for health professionals in its early stages, as it is difficult to identify it from medical imaging, which is an obstacle to starting adequate treatment for patients. This disease is the leading cause of death, with an increase in new cases, deaths and each year more people die from this cancer than from breast, prostate and colon cancer. Traditional classification techniques tend not to improve their assessment metrics due to their filtering, segmentation, feature extraction and classification processes. Traditional screening requires a large amount of time and financial resources. The methodology consists of six steps: it starts with prior research to review different studies. Then, a data set is selected. In the third step, the most prominent architectures are chosen to classify in relation to the ImageNet dataset. The fourth stage configures the models for training and validation. The fifth stage evaluates the resource consumption and performance of the models. Finally, a web application using the best performing architecture is created. After analyzing the selected architectures, percentage metrics of 97% or more were obtained. However, the tests revealed that the accuracy and precision metrics reached percentages of 95% and 91%, respectively. In conclusion, Efficientb4_DA achieves better results reaching an accuracy of 95.32%, a precision of 91.29%, a sensitivity of 89.84% and an F-score of 90.54%.

Keywords: Lung cancer, detection, classification, machine learning, evaluation metrics

I. INTRODUCCIÓN

1.1. Realidad problemática.

Un problema de salud internacional es el cáncer, el estado vigente de canceres alrededor del mundo sitúa al cáncer de pulmón en el 1er lugar del sexo masculino y ocupa el 3er puesto entre las mujeres [1]. El informe elaborado por la Organización Panamericana de la Salud OPS [2] menciona que dentro de 10 años se prevé un crecimiento en el número de casos de cáncer en las mujeres, el número de nuevos casos y de muertes se incrementará en un 50% en América Latina y el Caribe, además alrededor de 541.000 nuevos casos y 445.000 muertes por cáncer de pulmón en las Américas. Este tipo de cáncer es la primordial causa de muerte y cada año fallecen más personas por este cáncer que por cáncer de próstata, mama y colon, esto es ocasionado por el consumo de tabaco que representa el 80% de fallecimientos y el porcentaje restante en su mayoría por humo de segunda mano, contaminación ambiental entre otros [3].

Otros factores a considerar aparte de los mencionados son el radón, asbesto, cambios genéticos, exposición a químicos o productos de combustión [3] Esta enfermedad presenta síntomas en las personas que lo padecen como pérdida de apetito y peso, dificultad al respirar, tos constante, sensación de cansancio y debilidad. [3] Debido a unos ejemplares de este cáncer se suele presenta síndromes específicos (horner, vena cava superior y paraneoplásicos) que son originados por otras patologías [3]

Para los profesionales de la salud detectar el cáncer de pulmón en un período temprana a partir de imágenes médicas es complicado [4] Un diagnóstico temprano del cáncer de pulmón apoyaría en la ubicación del estadio del cáncer, selección del tratamiento y cuidado del paciente sin embargo los síntomas suelen aparecer cuando se encuentran en una etapa avanzada disminuyendo las esperanzas de vida de aquellos que padecen de esta enfermedad. [3]

Para la detección de cáncer actualmente se utiliza diversos métodos computaciones entre ellos el aprendizaje automático ML y técnicas de procesamiento de imágenes. Sin embargo, presentan dificultados como en el uso de aprendizaje profundo para el diagnóstico de los nódulos pulmonares en las imágenes de tomografía computarizada ya que posee menor precisión debido a la estructura redundante, falta de datos de entrenamiento y requiere demasiado tiempo. [5] Es importante mencionar que los diseños existentes de sistemas de detección y clasificación del cáncer de pulmón se basan en técnicas diseñadas a mano y sus resultados en términos de precisión y otras medidas de

rendimiento son limitados [6]

Asimismo, la detección de la existencia de cáncer de pulmón se puede realizar de diversas formas, como imágenes por resonancia magnética (IRM), radiografía y tomografía computarizada (TC) estas técnicas requieren mucho tiempo y recursos económicos elevados. No obstante, para la detección del cáncer de pulmón, la TC ofrece un menor coste, un tiempo de obtención de imágenes rápidas y a una mayor disponibilidad [7] Solo las tomografías computarizadas (TC) no pueden brindar una interpretación adecuada a los médicos expertos, es complejo encontrar la región de crecimiento celular a partir de las imágenes médicas [8]

La mayor parte de las técnicas de detección de cáncer computarizadas propuestas actualmente, fallan en el manejo de la mejor tasa de precisión debido a sus combinaciones de técnicas de filtrado, técnicas de segmentación y clasificadores [9] El proceso de identificación de cáncer de pulmón mediante redes neuronales convencionales no logra la mayor precisión y tampoco una predicción más rápida de las manchas de cáncer incluso antes de que los pacientes reconozcan los síntomas del cáncer [10]

Sim embargo, existen soluciones que han propuesto varios investigadores, es el caso de [5] publicaron un artículo en China, en el que proponen un nuevo método de diagnóstico basado en la Red Neural Convolucional de Transferencia Profunda (DTCNN) y la Máquina de Aprendizaje Extremo (ELM) para detectar los nódulos en imágenes de TC en un menor tiempo y mayor precisión.

Por otra parte, en Arabia Saudita se propuso un sistema de detección y clasificación automática del cáncer de pulmón (ALCDC) para diagnosticar si los tumores encontrados en los pulmones son malignos o benignos utilizando el modelo de red neuronal convolucional (CNN) por parte del investigador [6]

En Malasia [7] plantearon una técnica de procesamiento de imágenes para extraer características del cáncer de pulmón a partir de imágenes de tomografía computarizada proporcionado un menor costo, un tiempo de obtención de imágenes rápido y mayor disponibilidad.

Los investigadores [9] en la India, realizaron una Red neuronal de propagación hacia atrás de perceptrón multicapa (MLP-BPNN) basada en la extracción de características SIFT

(Transformación de características invariantes de escala) junto con Bolsa de palabras (BOW) para una adecuada detección computarizada del cáncer por medio del método mencionado.

Del mismo modo [10] en la India, se diseña una red neuronal recurrente basada en la atención (ARNN), para identificar el cáncer de pulmón en etapas iniciales ayudando en el proceso de recuperación de los pacientes.

Asimismo, [8] en la India proponen la red neuronal convolucional 3D (CNN) para la identificación del cáncer de pulmón a partir de las tomografías computarizadas (TC) del paciente

Acorde con lo mencionado, se propone desarrollar un sistema web para la clasificación automática del cáncer de pulmón mediante técnicas de procesamiento de imágenes y algoritmos de aprendizaje automático (ML) para la detección del cáncer de pulmón en imágenes de tomografías computarizadas (TC) ya que el diagnostico en etapas tempranas ayuda a la recuperación y tratamiento adecuado de los pacientes sirviendo de apoyo para los profesionales de la salud. Después de adquirir el conjunto de datos las TC se realizará una etapa de procesamiento, convertir las img mediante el procesamiento requerido por el modelo de aprendizaje posteriormente se seleccionan las arquitecturas destacadas de ML a implementar para realizar las pruebas se tendrá en cuenta los indicadores de rendimiento y consumo a través de la ficha de observación y apoyado por el administrador de tareas y selección del lenguaje de programación en el cual por medio sus librerías invocaremos a los métodos de exactitud, presión, recall y f, los artículos mencionados aportan a nuestra investigación a través de su metodología, resultados, discusiones, conclusiones y trabajos futuros puesto en esas sección se menciona el proceso de ejecución que siguieron otros investigadores y que permite el desarrollo de este trabajo.

[5] realizaron la investigación, Deep Transfer Convolutional Neural Network and Extreme Learning Machine for lung nodule diagnosis on CT images en China. Actualmente el diagnóstico de nódulos en imágenes de tomografía computarizada (TC) de pulmón basado en el aprendizaje profundo requiere mucho tiempo y es menos preciso debido a la estructura redundante y a la falta de datos de adecuados entrenamientos. Por ello se propone un nuevo método de diagnóstico basado en la Red Neural Convolucional de Transferencia Profunda (DTCNN) y la Máquina de Aprendizaje Extremo (ELM), consta de dos partes: el preprocesamiento de nódulos pulmonares en imágenes de TC y el diagnóstico de nódulos pulmonares basado en el DTCNN-ELM propuesto. Primero los

parches de región de interés (ROI) de nódulos de las imágenes de TC de pulmón se preprocesan a la dimensión de 64 × 64 utilizando el método de relleno cero (Zero Padding) mientras tanto, las imágenes grises de entrada (1 canal) se convierten en imágenes RGB (3 canales) por duplicado tres veces, que se ajustan a la capa de entrada de DTCNN. En la segunda etapa se construye el DTCNN, que consta de un modelo DCNN preentrenado y un modelo DCNN objetivo, donde el modelo DCNN preentrenado se utiliza para extraer características universales para la clasificación de imágenes comunes, y el modelo DCNN objetivo clasifica los nódulos de manera eficiente y precisa con la ayuda del DCNN previamente entrenado luego se construye el ELM y se determinan los parámetros, incluyendo el número de nodos ocultos. Entonces el ELM se combina con el DTCNN y se utiliza como clasificador. En el DTCNN-ELM, las muestras de formación se introducen en primer lugar en la estructura DTCNN para obtener las características representativas. Luego, estas características se combinan en características unidimensionales y se consideran las entradas del modelo ELM. Después de eso, se aplica una operación inversa generalizada para el entrenamiento ELM. Finalmente, en la etapa de prueba, las muestras se preparan y se ingresan en el modelo entrenado DTCNN-ELM para obtener los resultados finales del diagnóstico. Los resultados experimentales muestran que nuestro novedoso modelo DTCNN-ELM logró el rendimiento con una precisión del 94,57%, una sensibilidad del 93,69%, una especificidad del 95,15%, un área bajo la curva del operador receptor (AUC) del 94,94% y un tiempo de prueba por nódulo de 0,5 ms, que tiene los resultados más fiables en comparación con los métodos actuales del estado de la técnica. En conclusión, el método DTCNN-ELM propuesto alcanza una precisión del 94,57% en el conjunto de datos LIDC-IDRI y una precisión del 100% en el conjunto de datos FAHGMU.

[11] realizaron la investigación Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography en China. Los médicos profesionales identifican la enfermedad y determinan los estadios del cáncer mediante métodos tradiciones por lo que el desarrollo de nuevas técnicas no invasivas para predecir con precisión los subtipos histológicos en los nódulos de adenocarcinoma de pulmón detectados por tomografía computarizada (TC) es inexistente. Por ello proponen investigar las redes de aprendizaje profundo y radiómica para predecir la clasificación del subtipo histológico y la supervivencia del adenocarcinoma de pulmón diagnosticado a través de imágenes de TC, el método propuesto consta de la inclusión de pacientes, preparación del conjunto de datos, selección de métodos de aprendizaje profundo, exploraciones de Radiomics y Algoritmo de clasificador combinado radiómico profundo. Primero se inscribió retrospectivamente un conjunto de datos de 1222 pacientes con adenocarcinoma de

pulmón de tres instituciones médicas. Se obtuvieron las imágenes de TC preoperatorias anonimizadas y las etiquetas patológicas de hiperplasia adenomatosa atípica, adenocarcinoma in situ, adenocarcinoma mínimamente invasivo, adenocarcinoma invasivo (CAI) con cinco componentes predominantes. Estas etiquetas patológicas se dividieron en clasificación de 2 categorías (CAI; no CAI), 3 categorías y 8 categorías. El modelo de aprendizaje profundo se entrenó a partir de MedicalNet, se eligió 3D-ResNet-34 como la columna vertebral de este estudio, modelamos la tarea de clasificación de subtipos histológicos basados en la red de aprendizaje profundo ResNet-34 modificada, al mismo tiempo, exploramos el modelo radiómico para clasificar diversos subtipos de adenocarcinoma Luego, entrenamos un modelo de regresión logística multinomial penalizado con el operador de selección y contracción mínima absoluta (LASSO) utilizando el paquete glmnet. Este estudio incluyó un conjunto de entrenamiento (n = 802) y dos cohortes de validación (interna, n = 196; externa, n = 224). La precisión (ACC) del algoritmo deep radiomics en la validación interna alcanzó 0,8776, 0,8061 en la clasificación de 2 y 3 categorías, respectivamente. Incluso en la categoría 8, el área bajo la curva ROC (AUC) osciló entre 0,739 y 0,940 en el conjunto interno. Además, construimos un modelo de pronóstico cuyo índice C fue de 0,892 en el conjunto de validación interna. En conclusión, el sistema de triaje automatizado basado en radiómica profunda ha logrado un gran rendimiento en la clasificación de subtipos y la predictibilidad de la supervivencia en pacientes con nódulos de adenocarcinoma de pulmón detectados por TC, proporcionando la guía clínica para las estrategias de tratamiento.

[12] realizaron la investigación Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level CoOccurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods en Indonesia. Diagnosticar en una etapa temprana el cáncer de pulmón en función de los tipos benignos y malignos que se pueden ver en las imágenes de tomografía computarizada TC para acelerar el proceso de tratamiento. Por ello desarrollaron un sistema de detección de cáncer de pulmón basado en imágenes de tomografía axial computarizada (TAC), el diseño del sistema consta de 6 partes, la primera es la entrada de img de TC en formato jpg, el segundo proceso es el pre-procesamiento que consta de 2 partes la escala de grises para mejorar la calidad de la img y la conversión de las imágenes en escala de grises a binarias mediante el método de umbralización, la tercera parte segmentación toma los resultados obtenidos de la etapa anterior y luego se utiliza para tomar el área detectada por el cáncer de la imagen original, el sistema toma el valor más 1 píxel entre el área de la tomografía pulmonar pre-procesada, la cuarta extracción de características basada en la forma se realiza calculando el valor del área del

objeto canceroso segmentado que luego se reconstruye con el color de la imagen original una vez detectado el objeto canceroso, se calcula el área utilizando la función de área del contorno en OpenCV. La etapa de clasificación se lleva a cabo mediante el método Support Vector Machine (SVM) se realiza un proceso de entrenamiento y prueba, las entradas utilizadas en esta etapa son los valores de los parámetros en forma de área de cáncer, contraste, energía, entropía y homogeneidad mientras que la salida producida en la etapa de clasificación es una decisión en forma de normal, benigno o maligno y la última etapa es la toma de decisiones después de obtener los parámetros de la imagen de entrada las decisiones resultantes son las salidas producidas en la etapa de clasificación. La extracción de características Matric de co-ocurrencia de nivel de gris (GLCM) en la imagen producirá parámetros de contraste, energía, entropía y homogeneidad. En conclusión, el nivel de precisión basado en la decisión del sistema para determinar el diagnóstico de cáncer de pulmón benigno o maligno es del 83,33%.

[6] realizo la investigacion An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network en Arabia Saudita. Los diseños existentes de sistemas de detección y clasificación del cáncer de pulmón se basan en técnicas diseñadas a mano y sus resultados en términos de precisión y otras medidas de rendimiento son limitados. Por lo que se propone un sistema de detección y clasificación automática del cáncer de pulmón (ALCDC) para detectar y clasificar si los tumores encontrados en los pulmones son malignos o benignos utilizando el modelo de red neuronal convolucional (CNN), el método propuesto se divide en tres pasos. En el primer paso, se propone el método para Adquirir las imágenes de tomográfica computarizada (TC) de la base de datos LIDC- IDRI y después de una revisión se utilizaron 833 exámenes de TC ya que algunos contenían errores y otros eran muy pequeños (-3mm), después cuatro expertos llevaron a cabo el procedimiento de registro de nódulos en la base de datos. Cada experto examinó las pruebas individualmente luego los resultados fueron analizados otra vez por cuatro expertos, la clasificación benigna o maligna es logrado primero mediante la medición, es decir los nódulos malignos son aquellos casos con moderadamente sospechosos o muy sospechosos y los nódulos benignos muestran altos o signos modestos de un tumor benigno. En el segundo paso, la segmentación de los nódulos se lleva a cabo mediante marcas de expertos. Los datos se recopilaron de un archivo de Lenguaje de Marcado Extensible (XML) que contiene coordenadas de nódulos con los criterios analíticos de cada experto para segmentar nódulos. Hay 8296 nódulos (4329 benignos y 3967 malignos), cada corte de TC 2-D se usa luego como una muestra de entrenamiento. La imagen de TC se reduce a 28x28, para realizar el entrenamiento en la arquitectura CNN. Las imágenes

de dimensiones inferiores y dimensionales más grandes se tomaron y se redimensionaron en 28x28 imágenes de fondo luego, cada imagen de TC se ingresa para la clasificación de la CNN. Finalmente, tras la evaluación de las pruebas, el diagnóstico se completa con la red neuronal convolucional (CNN) en tejido benigno o maligno. Los resultados indican que el sistema ALCDC propuesto da una precisión del 97,2%. La comparación muestra que el sistema ALCDC propuesto funciona mejor que los sistemas de última generación existentes. El ALCDC propuesto será útil en la investigación de diagnósticos médicos y los sistemas de atención médica. En conclusión, se probó la base de datos LIDC-IDRI y los mejores resultados se obtuvieron con el 97,2% de precisión, 95,6% de sensibilidad y 96,1% de especificidad, que supera los resultados obtenidos con otras técnicas de aprendizaje.

[13] realizo la investigación Automated Detection of Lung Cancer Using CT Scan Images en Bangladesh. Detectar el área de cáncer de pulmón en etapas tempranas y alcanzar la mayor precisión para mejorar tasa de supervivencia del paciente. Para ello se propone un enfoque automatizado en el que se utilizan imágenes de tomografía computarizada (TC) para identificar el cáncer de pulmón en su etapa inicial el método propuesto consta de dos etapas el pre-procesamiento y el post-procesamiento, en la primera etapa la adquisición de las imgs se divide el conjunto de datos en 3 lotes con 26 img cada una, se almacenan en MATLAB y se muestran como una imagen en escala de grises RGB (0 es negro, 1 es blanco y 0 a 1 varían entre negro y blanco) después para procesar la img se pasan a formato HSV donde H es el tono, la porción de color del modelo de color y se expresa como un número de 0 a 360 grados. S indica saturación, la cantidad de gris en el color de 0 a 100%. V indica valor que expresa la intensidad del color de 0 a 100% donde 0 representa completamente negro y 100 es el más brillante. Luego en la mejora de la imagen se utilizó el estiramiento de contraste que aumenta el contraste de la imagen al desarrollar el rango de valores de gravedad de la imagen para abarcar el rango deseado de 0 a 1 y elimina la ambigüedad que puede aparecer en diferentes regiones de la imagen del conjunto de datos. Después la operación de filtro se realiza en la imagen para aumentar la suavidad, nitidez y realce de bordes. Pasando a la segmentación divide la imagen en partes en función de lo similar atributos para extraer características importantes de la imagen para su posterior análisis. Finalmente, en el post-procesamiento incluye la extracción de características e identificación, en la primera La operación region-propos de MATLAB proporciona una serie de propiedades para cada forma de la imagen. Nuestra investigación ha utilizado cuatro propiedades que incluyen área, circularidad (redondez y diámetro) y solidez. También aplicamos el algoritmo de selección de características de análisis de componentes principales (PCA) para verificar que las características seleccionadas sean

apropiadas para detectar la región de interés (ROI) cuando una región cae dentro de la longitud del valor de cada característica, entonces se identifica el ROI exitosamente. Nuestros resultados manifiestan una precisión del 96,15%, 92,30% y 96,15% para cada 3 lotes respectivamente además en la presente investigación no solo se ha identificado la región del cáncer de pulmón, sino también la no deseada se abordó el problema y hemos introducido otro enfoque para eliminar la región no deseada. En conclusión, este enfoque propuesto por el estudio fue capaz de lograr una tasa de precisión del 95% que es aceptable en la patología laboratorio.

[7] en la investigación, Classification of Lung Cancer Stages from CT Scan Images Using Image Processing and k-Nearest Neighbours en Malasia. La detección de la existencia de cáncer de pulmón se puede realizar de diversas formas, como imágenes por resonancia magnética (IRM), radiografía y tomografía computarizada (TC). Estas técnicas requieren mucho tiempo y recursos económicos. No obstante, para la detección del cáncer de pulmón, la TC ofrece un menor coste, un tiempo de obtención de imágenes rápido y una mayor disponibilidad. Por ello se propone una técnica de procesamiento de imágenes para extraer características del cáncer de pulmón a partir de imágenes de tomografía computarizada. Este proyecto se dividió en cuatro partes principales: primero la recopilación de datos y preprocesamiento, después la segmentación de imágenes, luego la extracción de características y culmina con el clasificador. La recopilación de datos fue la primera parte, donde se obtuvo las bases de datos del Instituto Médico y Dental Avanzado (AMDI), Universiti Sains Malaysia. Antes de que se recopilaran los datos, la ética fue aprobada por el Comité de Ética en Investigación en Humanos de USM (JEPeM) de la Facultad de Ciencias Médicas, USM, IPPT, Bertam, Pulau Pinang. A continuación, se formuló el marco de segmentación de la imagen, la imagen original de la base de datos fue segmentada manualmente mediante imagej y Adobe Photoshop CS6 software, se extrajeron las características de las imágenes como la medida dimensional física se midió como medida de la forma el área, el perímetro y el centroide fueron las únicas características que se consideraron para medir y la parte final El kNN se utilizó para clasificar el patrón de entrada de dos clases que fueron las etapas 1, 2, 3 y 4 después de que se extrajeron los datos de las características, kNN clasificó los datos en forma de vector. Los resultados muestran que kNN predijo la clasificación incorrecta con 37/2000 o 0,0185. Por tanto, la precisión del estudio fue del 98,15%. En conclusión, este estudio se llevó a cabo con éxito para clasificar las etapas del cáncer de pulmón utilizando kNN además se recomienda que el trabajo futuro debe centrarse en la recopilación de datos para poder clasificar las diferentes etapas del cáncer de pulmón.

[9] en su investigación, Efficient Computerized Lung Cancer Detection Using Bag of Words en India. Aunque muchas técnicas de detección de cáncer computarizadas se propusieron anteriormente, esas técnicas fallan en el manejo de la mejor tasa de precisión debido a sus combinaciones de técnicas de filtrado, técnicas de segmentación y clasificadores. Por ello se propone una red neuronal de propagación hacia atrás de perceptrón multicapa (MLP-BPNN) basada en la extracción de características SIFT (Transformación de características invariantes de escala) junto con Bolsa de palabras (BOW), en la metodología de trabajo primero se recopila un conjunto de datos de 300 imágenes pulmonares del Instituto y Centro de Investigación del Cáncer Rajiv Gandhi, Delhi, de las cuales 100 imágenes se utilizan para pruebas y 200 imágenes para entrenamiento las imágenes de escaneo tomografía computarizada (TC) se proporcionan como entrada que luego se procesa previamente donde se aplica la técnica de filtrado además se propone una técnica de filtrado High Boost para el preprocesamiento ya que la mayoría de las imágenes de TC contienen ruido que puede dar lugar a resultados inexactos y las características se extraen de la imagen filtrada donde se propone la técnica de extracción de características SIFT. Estas características extraídas se agrupan mediante agrupación de KMeans que luego se agrupan en palabras mediante el uso de BOW, es una técnica de clasificación especial que toma las características agrupadas y crea una palabra. Estas palabras se proporcionan como entrada a la Red neuronal de propagación hacia atrás basada en MLP donde las imágenes se entrenan y prueban. Los resultados se comparan en términos de precisión, especificidad, sensibilidad y los resultados se comparan con el sistema de detección de cáncer basado en filtros KNN y m3 y también con el sistema de detección de cáncer basado en SURF y transformada Wavelet, donde el sistema propuesto ofrece un mejor rango de precisión del 89% en comparación con otros disponibles que se logra utilizando el método BOW junto con la extracción de características SIFT y la técnica BPNN basada en MLP. En conclusión, el sistema propuesto es adecuado para la detección computarizada del cáncer de pulmón, ya que proporciona una mayor precisión en comparación con todos los sistemas existentes

[10] en su investigacion, Attention Based Recurrent Neural Network for Lung Cancer Detection en India. El proceso de detección de cáncer de pulmón mediante redes neuronales convencionales no logra la mayor precisión y tampoco una predicción más rápida de las manchas de cáncer incluso antes de que los pacientes reconozcan los síntomas del cáncer. Por lo que se diseña una red neuronal recurrente basada en la atención (ARNN) La metodología compuesta está compuesta por 3 etapas técnica de pre-

pocesamiento, extracción de características y la clasificación. Las entradas son las tomografías computarizadas (TC) de los pacientes pasando a la reducción de ruido, corrección de bordes y detección de la nitidez de la imagen mediante el filtro mediano. Luego en la extracción de características se inicia definiendo los parámetros, se calcula la distancia entre la muestra de prueba y todo el entrenamiento, ordenamos la distancia, usamos el k vecinos más cercanos (KNN) y reunimos sus categorías pasando a la aplicación de la mayoría simple de la categoría. Finalmente se armar un clasificador que separe entre vóxeles que tienen un lugar con tejido vascular típico y aquellos que tienen un lugar con territorios nodulares. Las comparaciones de precisión obtenida para el enfoque propuesto con los valores de la red neuronal convencional muestran que la tasa de precisión de nuestro enfoque propuesto para la imagen de entrada de la muestra 1 es alta (es decir, 99,3%) que la red neuronal tradicional (98,3%). En conclusión, nuestra arquitectura propuesta ha detectado cáncer de pulmón utilizando dos imágenes de muestra de entrada. Las manchas cancerosas se detectan con gran precisión (97,4%) y con valores de sensibilidad mejorados.

[14] en su investigación, Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques en Arabia saudita. Detección y clasificación de diferentes tipos de cánceres de pulmón como el adenocarcinoma, el carcinoma de células grandes y el carcinoma de células escamosas. Para ello se propone una nueva técnica de detección de cáncer de pulmón utilizando técnicas de aprendizaje automático. El marco de investigación propuesto consta de datos de entrada, pre-procesamiento, extracción de características, fusión y clasificación. En la primera las imágenes del conjunto de datos de la tomografía computarizada de tórax se han procesado previamente convirtiéndolas en escala de grises. Las imágenes de entrada se introducen en la técnica Patrón binario local (LBP) basada en texturas para codificar las características globales de las tomografías computarizadas de pulmón. Luego la selección de características de textura global implica tres etapas clave. En la primera sección describe que el dolor lumbar de tres parches se utiliza para la selección de características de textura de las imágenes de tomografía computarizada de tórax. Las segundas características se extraen a través de la Transformada de coseno discreta utilizando la etapa anterior y luego se combinan en la Segunda Sección. Finalmente, la clasificación en la Tercera Sección presenta clasificadores de aprendizaje automático como la máquina de vectores de soporte basada en el núcleo polinomial y el KNN para clasificar las imágenes de tomografía computarizada de tórax en cancerosas y no cancerosas. Los resultados de rendimiento muestran que la técnica propuesta logra un mejor rendimiento en el conjunto de datos de imágenes de

tomografía computarizada de tórax. En conclusión, la técnica propuesta logra una mejor eficiencia en términos de especificidad, sensibilidad y precisión promedio del 95%, 86% y 93%, en SVM, mientras que para KNN 93%, 82% y 91% de especificidad, sensibilidad y precisión, respectivamente.

[8]en su investigacion, Lung Cancer Detection using 3D Convolutional Neural Networks en India. Las tomografías computarizadas (TC) no pueden brindar una interpretación adecuada a los médicos expertos, es complejo encontrar la región de crecimiento celular a partir de las imágenes médicas, no obstante, se puede usar redes neuronales profundas, reconocimiento de patrones y clasificación de las imágenes para detertar el cáncer. Se propone usar la red neuronal convolucional 3D (CNN) para la identificación del cáncer de pulmón a partir de las tomografías computarizadas (TC) del paciente, el método propuesto contiene los siguientes pasos Cargar la imagen, Convertir a unidades Hounsfield, Extrae la máscara del nódulo, Aplicar la máscara de nódulo a la imagen original, Recortar, cambiar el tamaño y la escala de la imagen, Construir modelo de CNN, Realizar entrenamiento y validación y Evaluar el desempeño del modelo. Primero El conjunto de datos utilizado para entrenar a nuestra CNN es el SPIE-AAPM Lung CT Challenge del conjunto de datos de entrenamiento, hemos tomado 60%, 20% y 20% para entrenamiento, validación y pruebas, respectivamente. En la sección de preprocesamiento el primer paso es 1) Conversión a Unidad Hounsfield (HU): Extraemos la matriz de píxeles de los archivos DICOM del paciente y los clasificamos según el Número de instancia Luego, los convertimos a Hounsfield Unit (HU) con la ayuda de los valores de Rescale Slope y Rescale Intercept obtenidos de los metadatos de DICOM. En el siguiente paso 2) Umbral de la img: La porción pulmonar tiene una medida de radiodensidad específica de -500 HU. Hemos encontrado que el umbral de -420 HU se ajusta perfectamente a la segmentación pulmonar en nuestro caso. Después en la 3) Eliminación de las ubicaciones de los píxeles de aire sobre la bandeja del paciente: Hemos etiquetado cada región conectada (excepto las regiones cuyos valores de píxel son 0) con un valor diferente. Pasando a 4) Llenar el área del pulmón: Aquí primero hacemos una copia de la máscara, luego tomamos cada corte de la máscara copiada secundaria y convertimos sus valores de píxeles en binarios. Después 5) Eliminamos las bolsas de aire y aire debajo de la bandeja: ahora restamos 1 de cada valor de píxel, convirtiendo así la máscara en binario. Para 6) Obtener la máscara de nódulo: hemos generado dos máscaras diferentes, una con operación de llenado de área pulmonar y otra sin operación de llenado de área pulmonar. Luego, restamos la máscara pulmonar llena con la otra máscara, lo que da como resultado una nueva máscara con solo nódulos. 7) Aplicamos la máscara del nódulo a la imagen original: Esta operación conservará solo

los nódulos en la imagen original y 8) Recortamos la imagen del nódulo: en este paso se recortó 100 cortes del archivo de imagen DICOM, utilizando la ubicación del centro de la imagen del nódulo mencionada en la anotación del conjunto de datos. En el 9) Cambio de escala de la imagen del nódulo: cada corte en la imagen del nódulo se cambia a un espaciado uniforme de (1, 1, 1) utilizando el grosor del corte y los valores de espaciado de píxeles disponibles en los metadatos de los archivos DICOM mediante el método de interpolación spline. Para 10) Cambiar el tamaño de la imagen del nódulo: finalmente tenemos las imágenes de nódulos con un tamaño uniforme de 100 x 512 x 512. Entonces, para reducir la complejidad, cambiamos el tamaño de cada uno de los cortes en la imagen usando interpolación bilineal. Así, de esta manera, hemos convertido la imagen del nódulo en un tamaño de 100 x 100 x 100. Finalmente, en la fase experimental en la 1) Arquitectura CNN: creamos un modelo CNN 3D con dos capas convolucionales y dos capas completamente conectadas después el 2) Algoritmo de entrenamiento: entrena la CNN propuesta, hemos aplicado Adam optimizer, diseñado para entrenar una amplia gama de arquitecturas de aprendizaje profundo. Luego en el paso final 3) Los Hiperparámetros: juegan un papel esencial en la regulación del desempeño de un modelo una afinación adecuada de los hiperparámetros puede contribuir enormemente a optimizar un modelo. El resultado después de completar el proceso de entrenamiento arroja un 83,33% como precisión de entrenamiento, el modelo propuesto nos proporciona una excelente precisión de prueba del 100%. La razón para obtener una precisión del 100% podría deberse al pequeño conjunto de datos tiene una menor cantidad de instancias. En conclusión, el conjunto de datos SPIE-AAPM Lung CT Challenge es un conjunto de datos muy pequeño de aproximadamente 12 GB. Los alentadores resultados obtenidos aquí nos motivan mucho a probar el modelo 3D basado en CNN para la detección del cáncer de pulmón en conjuntos de datos más grandes.

[4] en su investigación, Lung Cancer Detection using CT Scan Images en India. Aunque la tomografía computarizada es la mejor técnica de imagen en el campo médico, es difícil para los médicos interpretar e identificar el cáncer a partir de imágenes de tomografía computarizada. El objetivo principal de esta investigación es evaluar las diversas técnicas asistidas por ordenador, analizando la mejor técnica actual y descubriendo sus limitaciones e inconvenientes y, finalmente, proponiendo el nuevo modelo con mejoras en el mejor modelo actual. El mejor modelo actual es el de Ignatious & Joseph “Computer aided lung cancer detection system.” En base a ello se agregaron mejoras al modelo actual y se propuso el nuevo modelo, en lugar del filtro de Gabor, se han implementado el filtro mediano y el filtro gaussiano en la etapa de pre procesamiento. Después del pre

procesamiento de la imagen procesada se segmenta utilizando la segmentación de la cuenca. Esto da la imagen con los nódulos de cáncer marcados. Además de características como el área, el perímetro y la excentricidad, características como el Centroide, el Diámetro y la Intensidad Media del Píxel han sido extraídas en la etapa de extracción de características para los nódulos de cáncer detectados. El mejor modelo termina después de la detección del nódulo canceroso, es la extracción de características y el cálculo de la precisión. Pero, su clasificación como benigno o maligno no ha sido implementada. Por lo tanto, se ha realizado una etapa adicional de clasificación de nódulos cancerígenos utilizando la máquina vectorial de apoyo. Las características extraídas se utilizan como características de entrenamiento y se genera un modelo de entrenamiento. Luego, el nódulo canceroso detectado desconocido se clasifica usando ese modelo de predicción entrenado. Comparando la precisión del modelo propuesto con el actual se puede ver que hay un aumento progresivo de precisión del 88,4% al 92%. La sensibilidad se mantuvo igual. La especificidad aumentó del 40% al 50%, además la clasificación del nódulo como maligno o benigno que no se realizó en el mejor modelo, en nuestro modelo clasifica como benigno o maligno con una precisión de 86,6%. En general, El modelo propuesto detecta el cáncer con una precisión del 92%, que es superior al modelo actual, y el clasificador tiene una precisión del 86,6%, podemos ver una mejora en el sistema propuesto en comparación con el mejor modelo actual. Sin embargo, este propuesto no clasifica en diferentes estadios como estadio I, II, III, IV del cáncer.

[15] en su investigación, Lung Cancer Detection Using Image Processing and Machine Learning HealthCare en India. Los médicos profesionales identifican la enfermedad y determinan los estadios del cáncer mediante procedimientos quirúrgicos, tratamiento químico, radioterapia y terapia dirigida los cuales son muy extensos, costosos y la parte del cuerpo afectada queda con dolor/ardura. Por lo tanto, para reducir este proceso se hace uso de varios algoritmos de procesamiento de imágenes. Se diseñó un sistema de detección de cáncer de pulmón y sus etapas utilizando procesamiento de imágenes y aprendizaje automático. Para poder detectar en cáncer de pulmón y sus etapas, en el primero se adquieren las imágenes de tomografía computarizada CT del usuario ya que en comparación con las img de rayos x y resonancia magnética RM la CT tiene menos ruido y se usa la basa de datos de ELCAP Public Lung Image DB, que contiene cerca de 200 imágenes pulmonares del informe de tomografía computarizada de pacientes cancerosos y no cancerosos, además de una muestra de sangre como entrada con imágenes de Tomografía axial computarizada TAC, después el pre procesamiento de imágenes se utiliza para mejorar la contradicción y claridad de las imágenes. Por lo tanto, se aplican varias

técnicas como la conversión de escala de grises (una imagen RGB se convierte en una imagen en escala de grises), la reducción de ruido (mecanismo de descartar el ruido de la imagen en escala de grises) y las técnicas de binarización (proceso de convertir la imagen en escala de grises en una imagen binaria) para obtener la imagen en la forma requerida. Luego el proceso de segmentación consta de algunos pasos. En primer lugar, transforma la imagen original / real en una imagen de solo borde. La imagen de solo borde transformada en imagen dilatada e imagen llena y, finalmente, ambos (pulmones izquierdo y derecho) se segmentan. Después la extracción de características es una etapa esencial que utiliza algoritmos y técnicas para reconocer los patrones de una réplica. La salida segmentada se proporciona como entrada para la extracción de características. Las siguientes características se tratan en la extracción de características como Área, Perímetro y Excentricidad y todas son de calidad escalar y finalmente el clasificador SVM divide la textura en dos grupos o clases, es decir, imágenes normales y anormales. Se utiliza para rastrear el nódulo con éxito. Los resultados indican que el método más eficiente para detectar el cáncer de pulmón y sus estadios con éxito es el uso de SVM y técnicas de procesamiento de imágenes. En nuestro sistema propuesto estamos describiendo el cáncer de pulmón y sus etapas usando diferentes algoritmos de procesamiento de imágenes y aprendizaje de máquinas como, conversión en escala de grises, reducción de ruido y binarización. Todos estos algoritmos se utilizan para el pre-procesamiento de la imagen de la tomografía computarizada dada.

[16] en su investigación, Deep Learning for Categorization of Lung Cancer CT Images en Filadelfia. La precisión deficiente en la detección temprana mediante la automatización del diagnóstico inicial de los escáneres médicos. Se desarrolló una red neuronal de convolución (CNN) y una tubería de reprocesamiento para aumentar la precisión del proceso de selección inicial. Se usa la BD de imágenes de cáncer de pulmón de “Kaggle Data Science Bowl 2017”, contando con más de 1500 instancias, cada paciente posee una etiqueta positiva o negativa para el cáncer. El grosor de corte, calidad de imagen, parámetros de exploración varían de un paciente a otro, el método posee 3 etapas: Pre-procesamiento, Redes neuronales convolucionales CNNs y predicción. En el primero se usó las img sin procesar (escaneo medio) y las etiquetas se importan y pre procesan y se crean conjuntos de pruebas y entrenamiento aleatorio como entradas para las CNNs, después se aplica el filtro gaussiano en las imágenes de prueba y entrenamiento para crear un segundo conjunto de pruebas de entrenamiento suavizado. Es decir, las img suavizadas y sin suavizar se parchean, mostrándolas en pequeños subconjuntos superpuestos y recién se utilizan como entrada para la convolución y agrupación. En la etapa 2 los datos de entrada se introducen en cada una de las CNN y esta las procesa usando un filtro de

convolución (10x10), luego las funciones convolucionales pasan a la capa de agrupación máxima, las características convolucionales y agrupadas de cada CNN son ingresadas en sus respectivas capas “Softmax” y completamente conectadas esto completa cada CNN y en esta última fase se implementa el mecanismo de votación, el sistema de votación funciona dando el resultado acordado por las dos CNN o dando un resultado negativo. Los resultados iniciales de nuestro mejor método muestran una alta precisión (97,5%) y un bajo porcentaje de falsos positivos (<10%). En conclusión, las tasas de falsos positivos en el diagnóstico clínico real llegan a un 95% sin embargo podemos ver que nuestro simple método de votación no aumenta significativamente nuestra precisión. Esto es más probable porque el método de votación apunta específicamente los falsos positivos que ya son bajos para esta implementación.

[17] en su investigación, Lung Cancer Classification Using Deep Learned Features on Low Population Dataset en India. Insuficiencia de muestras para aprender un mapeo preciso entre las características y las etiquetas clase. Se propone un mecanismo de clasificación de autocodificadores profundos que primero aprende características profundas y luego entrena una red neuronal artificial con estas características aprendidas. El método se divide en dos etapas, el propósito de la primera etapa es escoger clasificadores que se desempeñan relativamente mejor en este conjunto de datos para poder examinar con más detalle los mejores resultados. La segunda/última etapa utiliza estos clasificadores y utiliza la validación cruzada estrategia para asegurar la comparación de tarifas entre clasificadores. En la etapa inicial, los datos se dividen en formación y de pruebas para asegurar un resultado imparcial de la prueba. 24 muestras de los datos originales (75%) se utilizan para el entrenamiento, mientras que el resto las muestras se conservan para la validación de la retención. La división es hecha de acuerdo con el muestreo estratificado en las etiquetas de clase para asegurar El conjunto de capacitación y pruebas contiene un porcentaje similar de muestras de cada clase. Todos los clasificadores del experimento inicial se entrenan con este 75% de los datos de entrenamiento. Una vez que los clasificadores se entrenan, se aplican en el set de espera y su el rendimiento se registra. En la segunda o última etapa, los clasificadores de mejor desempeño de etapa inicial se seleccionan para una investigación a fondo utilizando validación cruzada con múltiples ejecuciones, con reentrenamiento y evaluación, para asegurar la integridad estadística. Esta validación cruzada de múltiples carreras estrategia es equivalente a la validación cruzada repetida que a menudo proporciona un rendimiento más estabilizado la medición de un clasificador. Los resultados experimentales muestran que el clasificador de aprendizaje profundo supera a todos los demás clasificadores cuando se entrenan con

todos los atributos y las mismas muestras de entrenamiento también se demuestra que la mejora del rendimiento es estadísticamente significativa. En conclusión, la aplicación del aprendizaje profundo tiene la potencial para aumentar significativamente la precisión de la clasificación para el conjunto de datos de cáncer de pulmón de baja población y alta dimensión sin requerir ninguna característica específica del caso, hecha a mano.

[18] en su investigación, A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks en Malasia. La precisión de los sistemas basados en CNN es menor para los sistemas de detección de cáncer de pulmón. El sistema propuesto tiene como objetivo aumentar el rendimiento de de la CNN en la clasificación de tumores malignos y benignos en imágenes de TAC para una mejor y temprana detección de probabilidades de cáncer. En Procesamiento de imágenes, las imágenes RGB se convertirán a escala de grises y binarias. Las imágenes se mejorarán y el ruido se eliminará mediante filtros. El efecto borroso se eliminará si lo hay. Esto mejora la calidad de la imagen de entrada. La caja de herramientas de procesamiento de imágenes en MATLAB se utiliza para realizar las etapas de procesamiento de imágenes. Son posibles muchos algoritmos diferentes para cada etapa del procesamiento de imágenes. El proceso de extracción de características en Convolution Neural Networks es tal que las características son definidas y calculadas por el propio algoritmo. Durante la etapa de entrenamiento, se proporcionan una etiqueta de entrada y salida. En base a los datos proporcionados, el algoritmo analiza las características / patrones y, para los datos de entrenamiento, forma un conjunto de parámetros y extracción de características. Según los cálculos, los nuevos datos se pueden probar para predecir una salida correcta. Las redes neuronales de convolución constan de una capa de entrada y una de salida, y varias capas ocultas. Las capas de entrada aceptan entradas y el número de capas de salida define el número de salidas en el resultado. Las capas de convolución se utilizan para definir características y parámetros. La agrupación de capas reúne los cálculos con una permutación similar. El filtro de convolución formará una salida espacialmente densa asignando un valor común a un conjunto de píxeles de la matriz. Estos valores deciden la salida de esa imagen. Resultados, se tomaron un total de 910 imágenes de LIDC se probaron un total de 281 imágenes, 47 imágenes de normal, 121 imágenes de benigno y 113 imágenes de clase maligna. Se obtiene una precisión general del 94,34% del sistema propuesto. En conclusión, el conjunto de datos LIDC es un conjunto de datos estáticos. Mantener una base de datos dinámica en tiempo real ayudará a estudiar los cambios en los casos de cáncer de pulmón durante un período de tiempo. Aumentar el número de convoluciones en el aprendizaje profundo mejorará los resultados del sistema de forma variable, pero también afectará exponencialmente el tiempo de retardo de

rendimiento y eficiencia.

La presente investigación está enmarcada en la línea de investigación de infraestructura, tecnología y medio ambiente dentro del área prioritaria de ciencias de la computación de la escuela académico profesional de ingeniería de sistemas de la universidad señor de Sipán, el trabajo es pertinente ya que existen diversas metodologías de procesamiento de imágenes y algoritmos de aprendizaje automático para realizar la detección y clasificación del cáncer de pulmón en etapas tempranas, se pretende corroborar si mediante estas dos tecnologías se podrá detectar el cáncer de pulmón en imágenes tomografías y mejorar el proceso de rehabilitación y tratamiento de los pacientes, sirviendo de apoyo a los médicos expertos, al mismo tiempo estaremos dando un aporte al conocimiento científico para que otros investigadores o desarrolladores puedan aplicar de manera precisa los resultados de la investigación en el desarrollo tecnológico. Desde el panorama tecnológico no existe ninguna restricción en la implantación del sistema propuesto ya que se ha desarrollado en diversos entornos de desarrollo y lenguajes de programación que hacen uso de sus librerías para invocar los métodos previamente establecidos. Económicamente en la etapa de desarrollo y pruebas en los artículos se hace mención igualmente al sistema operativo (SO), procesador y memoria RAM para ello este trabajo cuenta SO Windows, Core i7-7500U de Intel y 12GB que es equivalente a los equipos informáticos usados para evaluar el rendimiento de los indicadores. Además, se tiene acceso a la base de dato de la universidad donde se encuentran alojados los artículos científicos que se tomaron como base para aplicar nuestro trabajo de investigación.

1.2. Formulación del problema

¿De qué manera se podrá clasificar el cáncer de pulmón en imágenes tomográficas?

1.3. Hipótesis

Si se utiliza el procesamiento de imágenes digitales y aprendizaje automático entonces se podrá clasificar el cáncer de pulmón en imágenes tomográficas

1.4. Objetivos

Objetivo general

Clasificar el cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes digitales y aprendizaje automático.

Objetivos específicos

- Adquirir el data set de cáncer de pulmón
- Seleccionar las Arquitecturas de Aprendizaje automático a implementar

- Implementar las técnicas de Aprendizaje automático para clasificar el cáncer de pulmón
- Realizar pruebas sobre las técnicas implementadas
- Desarrollar una aplicación web

1.5. Teorías relacionadas al tema

1.5.1. Procesamiento de imágenes digitales

Debido a las deficiencias en las capturas de las imágenes surge este proceso que implica la edición de valores de los pixeles a fin de aumentar el contraste entre los objetos cuantificados y el fondo [19] Es decir, es un conjunto de tecnologías aplicadas a imágenes digitales para optimizar la calidad o la indagación de información, estos procesos incluyen: Conversión de imágenes, Corrección de defectos, Deconvolución, Ecualización, Filtros, Manipulación de contraste, Procesos matemáticos, Transformaciones geométricas [19]

1.5.2. Técnicas de procesamiento de imágenes digitales

Para la mayoría de investigaciones se pretende lograr el mejor contraste de la imagen antes que la perfección de ella, el realce para corregir fallas de iluminación debido a la abundancia o carencia de luz, contraste o equipos electrónicos además el procesamiento permite identificar con mayor facilidad al objeto y el fondo, el proceso de filtrado se puede llevar de 2 maneras en el domino de frecuencia y espacial [19]

Procesamiento en el dominio espacial

Transformaciones similares, afines y proyectivas: El primer ejemplo sobre transformación de dominio espacial es las transformaciones geométricas que son resultado de escalamientos, rotaciones y traslación de img de 2 dimensiones, en adelante para la explicar la siguiente sección se tendrá en cuenta las representaciones tradicionales de las imágenes donde: “y” numera los pixeles de superior a inferior (eje vertical), “x” numera los pixeles de lado a lado (eje horizontal) además de la ecuación de ángulo positivos medido en sentido antihorario, la ecuación 1 representa la ubicación X de un píxel, desarrollada en ejes homogéneos X_h. [20]

$$\mathrm{x} = \begin{bmatrix} x \\ y \end{bmatrix} \Longrightarrow \mathrm{X}_h = \begin{bmatrix} \alpha x \\ \alpha y \\ \alpha \end{bmatrix} \qquad (1)$$

El símbolo α es frecuentemente seleccionado como uno y diferente

de cero, y se diferencian las demás transformaciones básicas. [20]

$$\mathrm{u} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\theta & sen\ \theta \\ -sen\,\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = R_X \qquad (2)$$

La ecuación 2 representa un Giro en un Angulo θ en la superficie de la imagen partiendo del origen [20]

$$\mathrm{u} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = Z_X X \qquad (3)$$

La ecuación 3 es el Escalamiento isotrópico. [20]

$$u_h = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_X \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} I & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = T_{(t_x,t_y)h} X_h \qquad (4)$$

La ecuación 4 es la Traslación en coordenadas homogéneas[20]

- **Transformación euclidiana:** Caracterizada por la rotación y traslación, es relevante mencionar que no se debe de alterar el orden no de esta transformación, la matriz se obtiene del producto de $R_{\theta \mathrm{h}}$ con $T_{(\mathrm{tx,ty})\mathrm{h}}$ además posee tiene la característica de conservar distancias y espacios en las imágenes no modificables, la ecuación 5 representa la Matriz de trasformación euclidiana[20]

$$E_h = \begin{bmatrix} \cos\theta & \mathrm{sen}\,\theta & t_X \\ -\mathrm{sen}\,\theta & \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} R_X & t \\ 0^T & 1 \end{bmatrix} \qquad (5)$$

- **Transformación de similitud:** Es un tipo de transformación que conserva la forma de un objeto sin cambiar el tamaño [20]

$$S_h = \begin{bmatrix} s\cos\theta & s\,\mathrm{sen}\,\theta & t_X \\ -s\,\mathrm{sen}\,\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} sR_\theta & t \\ 0^T & 1 \end{bmatrix} \qquad (6)$$

  La ecuación 6 es la Matriz Homogénea, la transformación mantiene constante distancia entre la longitud y ángulo, también la matriz posee 4 categorías de libertad por la dependencia del parámetro θ, t_x, t_y y s s [20]

- **Transformación afín:** La conversión esta explicada por la matriz semejante llamada Matriz Homogénea y Submatriz en la ecuación 7 [20]

$$A_h = \begin{bmatrix} a_{11} & a_{12} & a_x \\ a_{21} & a_{22} & a_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} A & t \\ 0^T & 1 \end{bmatrix} \qquad (7)$$

A_h posee seis categorías de libertad, manteniendo paralelas las líneas y distancia entre longitudes de fragmentos de líneas paralelas entre ellas [20]

"La submatriz no debe de ser única, y puede descomponerse como se expresa en la imagen, es decir A es la matriz que representa el inicio de la rotación en un ángulo, después le sigue el escalamiento anisotrópico la cual distorsiona la img original a la escala "x" por un factor y en "y" por otro" [20]

$$A = R_\theta R_{-\phi} D_{(\lambda_1,\lambda_2)} R_\phi \quad (8)$$

La ecuación 8 es la Matriz de escalamiento anisotrópico [20]

- **Transformación Proyectiva:** Posee como forma general la siguiente matriz, debido a la transformación de homogénea a euclidianas sucede una normalización que muestra las categorías de libertad (8), no como el n° de variables (9) implicadas, la ecuación 9 representa la Matriz de transformación proyectiva. [20]

$$H_h = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ v_1 & v_2 & v \end{bmatrix} = \begin{bmatrix} A & t \\ v^T & v \end{bmatrix} \quad (9)$$

Transformaciones de niveles de gris:

"La conversión de grises es la aplicación básica de la manipulación de valores de píxeles. Utilizando la notación de imagen funcional, la imagen $f(x,y)$ se convierte en la imagen $g(x,y)$, donde T es la función escalar de la situación en cuestión" [20]

$$g(x,y) = T\big(f(x,y)\big) \quad (10)$$

La ecuación 10 es la Notación funcional de imágenes [20]

Teoría de sistemas y filtrado espacial: Las funciones de filtro se fundamentan en la teoría de identificación, simulación y reconstrucción de sistemas [20] Dentro de este apartado se encuentran:

- **Linealidad e Invarianza a la traslación:** Acorde con sistema T que convierte una img de ingreso a en una de

salida b, la ecuación 11 es el Sistema de transformación de img [20]

$$b(x,y) = T\,[a(x,y)] \tag{11}$$

se le nombra lineal si:

$$T[k_1 a_1(x,y) + k_2 a_2(x,y)] = k_1 T[a_1(x,y)] + k_2 T[a_2(x,y)] \tag{12}$$

La ecuación 12 es el Sistema Lineal y se considera invariante si:

$$T[a(x - x_0\,, y - y_0)] = b(x - x_0, y - y_0) \tag{13}$$

La ecuación 13 representa la Invariante a la translación, además, puede ser un sistema lineal e invariante al cambio si cumple con ambas propiedades [20]

- **Convolución**: Primeramente, se establece el valor semejante al impulso unitario empleado en el sistema de tiempo prudente unidimensional, que es aplicable a las img en 2 dimensiones, la ecuación 14 es el Impulso espacial unitario [20]

$$\delta(x,y) = \delta(x)\delta(y) = \begin{cases} 1 & si\ x = y = 0 \\ 0 & en\ el\ resto \end{cases} \tag{14}$$

Conforme con la función descrita en ecuación 15 “Función de una imagen sobre rejilla discreta” de una rejilla prudente es viable sacar el valor de los pixeles en las codenas “x” e “y” [20] A través de:

$$f(x,y) = f(u,v)\delta(x - u, y - v) \tag{15}$$

La multiplicación de las variables “u” e “v” es cero, menos cuando “u” es igual a “x” y “v” a “y”, estos pasos permitirán rehacer la imagen como suma de sus pixeles [20]

- **Filtros lineales:** existe variedad de filtros por ello en esta sección se mencionan los las usuales:

  Filtros de media móvil: Son conocidos como filtros promediados en la los valores son 1/S en la máscara cuadrada en la que “S” representa la cantidad de pixeles [20]

Aproximación del gradiente: Es una operación elemental para revelar esquinas, curvas, bordes, líneas entre otros, en las zonas de variación de las regiones semejantes se halla la información de las imágenes [20]

Derivaciones orientadas de la función gaussiana: La derivada recalca el murmullo se sugiere emplear un filtro de suavizado anteriormente de ello [20]

Kernels de Ando: Llamados también como Kernel optimizado ya que debido a ellos se puede mejorar magnitud, invarianza a rotación, función interpoladora, entre otras [20]

- **Filtros de rango:** Siendo un filtro no lineal, en la que no hay convolución, se sacan los pixeles sobre la mascarilla eligiendo uno acorde con su lugar en el arreglo de los pixeles [20]

Procesamiento en el dominio de la frecuencia

Análisis discreto de Fourier

- **Serie generalizada de Fourier:** Es una extensión de la serie de Fourier que permite la representación de funciones en intervalos arbitrarios. [20]
- **Serie discreta de Fourier:** Es un tipo de serie de Fourier que utiliza un número finito de términos para aproximar una función periódica

$$E_k(n) = e^{j2\pi kn/N}, k = 0 \dots N-1 \quad (16)$$

La ecuación 16 son los Exponenciales complejos armónicamente relacionados. "Entre ellos, cuando hay una frecuencia infinita correspondiente a un múltiplo entero infinito del período 2π / ω0, k puede crecer indefinidamente" [20]

- **Transformada de Fourier de señales discretas**

$$X(w) = \sum_{n=-\infty}^{\infty} x(n)e^{-jwn} \quad (17)$$

La ecuación 17 es la Transformada de Fourier de señal de energía limitada de tiempo prudente. "Donde X (ω) es igual a X (2πk), lo que significa que la banda de frecuencia única tiene un límite de [0, 2π [o equivalente] -, π], que diferencia en el rango de señal perenne" [20]

- **El teorema del muestreo:** En la siguiente figura “X (ω) es periódica con período base 2π, obtenemos que la frecuencia mínima de muestreo Fm debe elegirse de manera que sea al menos el doble de la frecuencia máxima de la señal” [20]

$$X(n) \circ\!\!-\!\!\circ X(w) \sum_{n=-\infty}^{\infty} x(n)e^{-jwn}, \quad w \in]-\pi, \pi] \qquad (18)$$

La ecuación 18 es la Representación de Fourier de la señal prudente. [20]

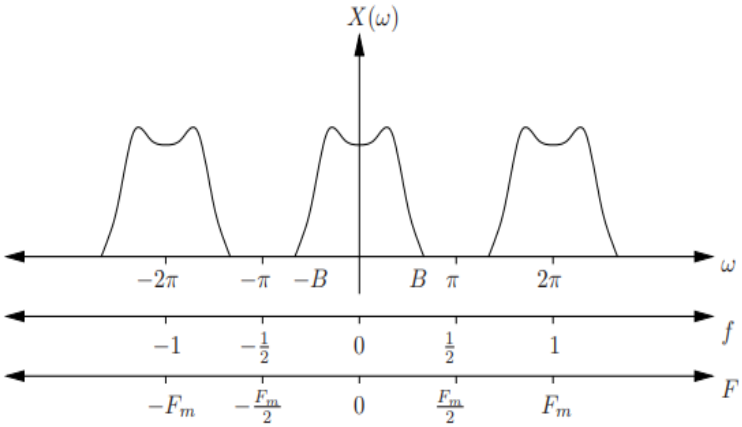


Figura 1. Espectro de señal independiente con rango de frecuencia limitado.

- **Propiedades de la transformada de Fourier de señales discretas:**

  Se caracterizan por Desplazamiento espacial, Desplazamiento frecuencial, Diferencia en el dominio de la frecuencia, Linealidad, Reflexión espacial, Simetría, Teorema de la Convolución, Teorema de la Correlación, Teorema de modulación, Teorema de Parseval yTeorema del enventanado [20]

- **Muestreo en el dominio de la frecuencia:** Es un proceso en el que se convierte una señal de tiempo en una señal de frecuencia. Esto se realiza mediante la aplicación de técnicas de procesamiento de señales digitales, como la Transformada Rápida de Fourier (FFT) [20]

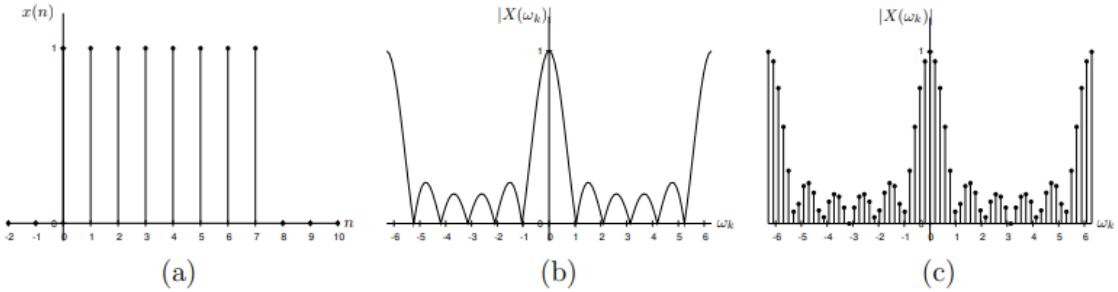


Figura 2. Espectro de señal prudente con banda establecida sin aliasing.

- **La transformada discreta de Fourier:** "En teoría, esto implica un procesamiento en el dominio de la frecuencia, el espectro dado a partir de una señal $x(n)$ de longitud finita L no está representado directamente por $x(n)$ sino durante una semana. Su retorno $x_p(n)$ se obtiene de:" [20]

$$x_p(n) = \sum_{l=-\infty}^{\infty} x(n - lN) \qquad (19)$$

La ecuación 19 es la transformada discreta de Fourier [20]

- **Relación de la DFT con otras transformadas:** Tiene correspondencia principalmente con las series de Fourier, las transformada de Fourier de secuencias aperiódicas y transformada z [20]
- **Propiedades de la DFT:** Las más empleadas son linealidad, simetría circular y convolución circulas, aunque existen otras las cuales se mencionan en el siguiente grafico [20]

1.5.3. Aprendizaje automático

El aprendizaje automático ML es uno de los campos de la Inteligencia artificial AI, mencionar que la segunda busca crear máquinas capaces de simular comportamientos inteligentes y ML ofrece un cambio de paradigma porque

busca dar a las máquinas la capacidad de aprender es decir se puede programar una máquina para moverse pero es diferente codificar una para que aprenda a moverse además este tipo de aprendizaje está dividido principalmente en tres grupos, aunque algunos consideras otros [21]

Los tipos de aprendizaje automático son: supervisado, no supervisado, reforzado, profundo y refuerzo profundo los cuales representan la evolución constate que ha tenido este campo comenzado por el Aprendizaje supervisado que surge a partir de enseñarle que resultados queremos obtener para un valor en específico, en contraste con el no supervisado el cual aprenderá por su cuenta sin ninguna supervisión, después se comenzó a considerar recompensar a las maquinas cuando cumple la tarea de forma esperada y nació el aprendizaje por refuerzo, luego la cantidad de datos sobrepaso a las técnicas tradicionales que no lograban el análisis de big data ni proporcionar predicciones de esta manera llega las redes neuronales ANN, actualmente el aprendizaje profundo ha logrado resolver gran parte de los problemas complejos que no tenían solución así mismo como se mencionó anteriormente un aprendizaje por refuerzo se otorgó incentivos a las redes neuronales con premios y finalmente llego el aprendizaje por refuerzo profundo [21]

Aprendizaje supervisado: [22] menciona que esta categoría se entrena a partir de un conjunto de datos en la que los resultados de salidas son conocidas, realizan ajustes en sus parámetros para adaptarse a los datos de entrada, después del entrenamiento adecuado el modelo realizara predicciones acertadas ante nuevos datos no procesados previamente.

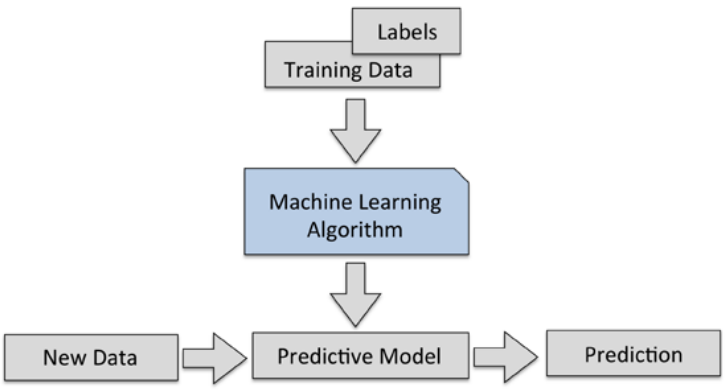


Figura 3. Aprendizaje Supervisado.

En el grafico se aprecia el proceso de este tipo de aprendizaje después del entrenamiento adecuado y aplicación del algoritmo se ingresan nuevos datos que no han sido visto por el modelo y finalmente logra predecir de manera correcta o errónea, las dos principales aplicaciones son la clasificación que tiene el objetivo de predecir las clases de una categoría por ejemplo detectar a un mensaje de spam que es una clasificación binaria es decir es spam 1 y no lo es 0. La regresión es usada en la asignación de categorías para datos sin etiquetar, se trata de buscar la relación entre dichas variables y obtener un resultado continuo [22]

Aprendizaje no supervisado: En este tipo de aprendizaje se caracteriza por la generación de conocimiento a partir de datos de entrada sin necesidad de explicarle al sistema que se quiere obtener, aquí existen dos subcategorías relevantes el agrupamiento o también llamado clustering que organiza la información similar que se diferencie de los demás objetos en conjuntos obteniendo información similar por grupos y la reducción dimensional es aquella que identifica relaciones entre los particularidades [22]

Aprendizaje por refuerzo: Se diferencia de su antecesor porque existe la necesidad de etiquetar los datos en entrada ni de salida centrándose en aprender nuevas soluciones y explotar aquellas entendidas, poniendo en contexto de caso del entrenamiento de animales, si a un delfín en enseñarle a saltar a través de un aro cada vez que lo haga recibirá una recompensa en este caso un pez y de manera lenta comprenderá que recibe algo a cambio de ejecutar una acción de manera correcta [21]

Aprendizaje profundo: Es conocido como Deep Learning está inspirada en la estructura neuronal de cerebro, usa una estructura jerárquica de redes neuronales artificiales ANN, es decir se aprende por niveles en las primeras capas se aprenden conceptos muy específicos como que es una rueda, un tornillo, un espejo, un asiento y en las capas posteriores se usa la información aprendida previamente para aprender conceptos más abstractos como un coche, un camión, una moto [22]

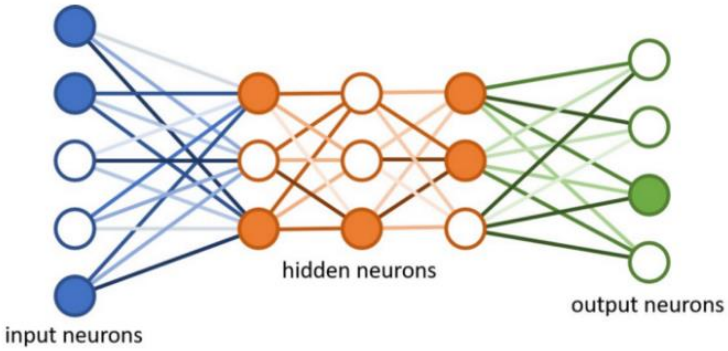


Figura 4. Red neuronal.

Aprendizaje por refuerzo profundo: En esta rama del machine learning el objetivo es la construcción de un modelo que mejora su rendimiento a partir de la recompensa que se obtiene en cada iteración realizada, un ejemplo clave para esto sería el ajedrez debido a que el agente decidirá entre múltiples acciones, dependiendo del entorno, la recompensa se da al culminar la partida y obtener resultados [22]

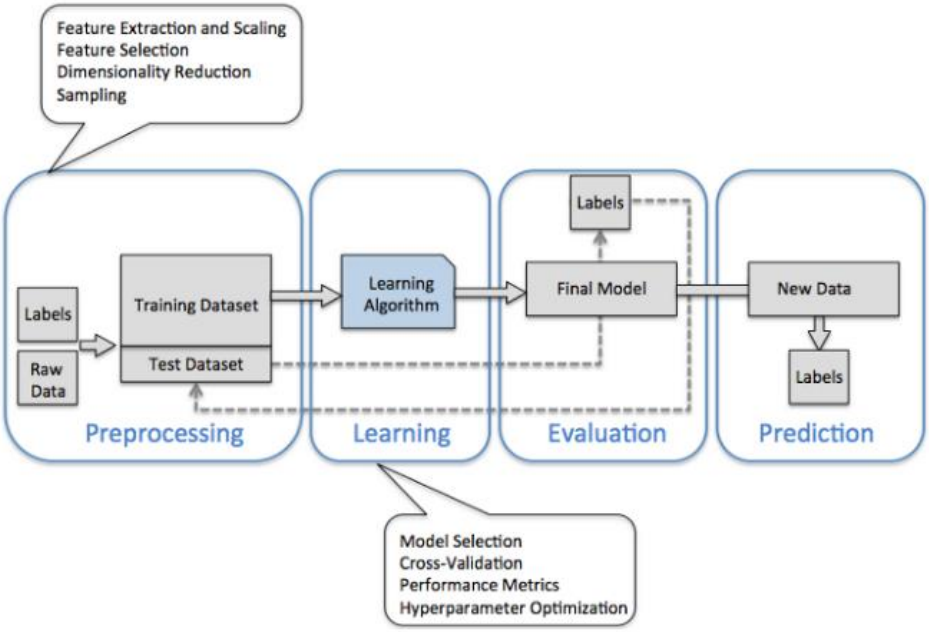


Figura 5. Método general para la construcción de un modelo de ML.

La imagen representa una metodología general contando de cuatro etapas, preprocesamiento en esta etapa los datos no incorrectos, por lo que se deberá de aplicar técnicas dependiendo del conjunto de datos después de ello en forma aleatoria se deberá de partir los datos en 3 entrenamiento,

validación y prueba ahora en la fase de entrenamiento y selección del modelo se comienza a identificar las métricas de evaluación para nuestro algoritmo finalmente en la evaluación del modelo se estima el rendimiento en base al ingreso de nuevos datos [22]

1.5.4. Algoritmos de aprendizaje automático

En este apartado se mencionarán los algoritmos de las categorías mencionadas anteriormente en el caso del aprendizaje supervisado los más empleados son:

- **k-Vecinos más cercanos:** Esta técnica estadística se emplea en la resolución de problemas de clasificación y regresión también son conocidos como KNN [21] Se caracteriza por memorizar los ejemplos de formación que después se usa como conocimiento en la fase de predicción [23]

  Si se selecciona como lenguaje Python tendiéramos que utilizar la librería Scikit Learn la cual cuenta con el modelo de implementación del algoritmo, después de la definición pasa a importar la clase a KNeighborsClassifier para este ejemplo de clasificación [23]

  La librería ofrece varios parámetros que tendrán que configurarse para mejorar el modelo que se está construyendo, el número de vecinos por defecto es 5 pero dependerá de cada desarrollados definir otro valor, después establecer la distancia euclidiana que se empleara en la verificación de los vecinos del dato que se busca predecir, el primer valor es p y el segundo metric, definiendo sus valores se podrá logar un modelo adecuado empleando KNN [23]

p : *integer, optional (default = 2)*

Power parameter for the Minkowski metric. When p = 1, this is equivalent to using manhattan_-distance (l1), and euclidean_distance (l2) for p = 2. For arbitrary p, minkowski_distance (l_p) is used.

metric : *string or callable, default 'minkowski'*

the distance metric to use for the tree. The default metric is minkowski, and with p=2 is equivalent to the standard Euclidean metric. See the documentation of the DistanceMetric class for a list of available metrics.

Figura 6. Parámetros de KNN.

En la etapa de entrenamiento y predicción del algoritmo se invoca a los

métodos preestablecidos al igual que el resto de algoritmos de Ml se llama a fit y predic y antes de ello se tendrá que definir los valores de x e y una vez culminado este apartado se procede a evaluar el rendimiento del modelo construido esto es relevante ya que se tendrá que comparar cómo funciona con otros algoritmos en conjunto de datos [23]

```
from sklearn.neighbors import KNeighborsClassifier
x_entrenamiento = variablesIndependientes_entrenamiento
y_entrenamiento = variableDependiente_entrenamiento
x_prueba = variablesIndependientes_prueba
y_prueba = variableDependiente_prueba

algoritmo = KNeighborsClassifier()
algoritmo.fit(x_entrenamiento, y_entrenamiento)
algoritmo.predict(x_prueba)
```

El código muestra un ejemplo de KNN de clasificación. [23]

- **Máquinas de vectores de soporte SVM:** "Es un clasificador distinto formalmente identificado por el Nivel Superior Separado" [23] Es decir, partir del conjunto de datos de entrenamiento este algoritmo genera una línea que separa a los ejemplos en 2 partes.

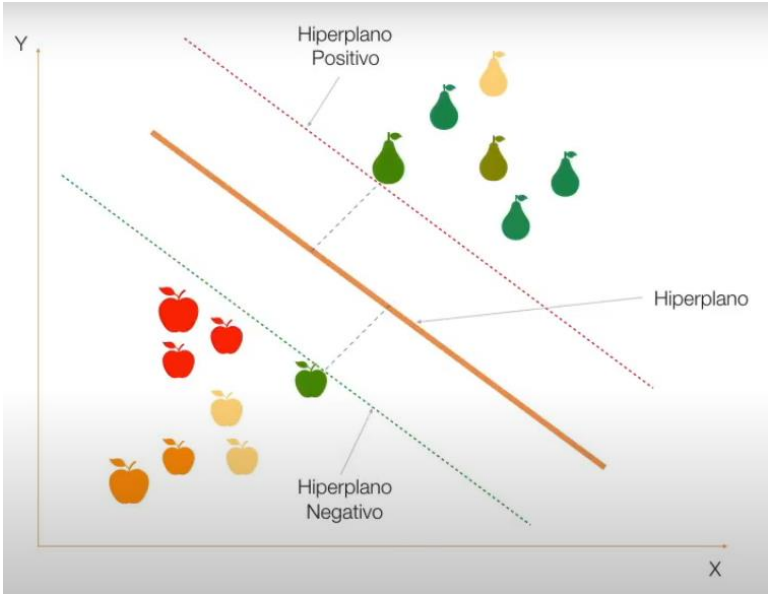


Figura 7. Ejemplo de SVM de clasificación.

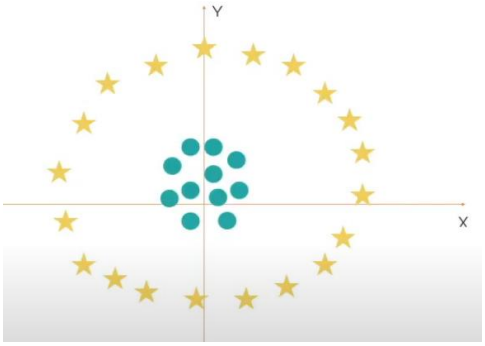


Figura 8. Ejemplo de SVM no lineal parte 1.

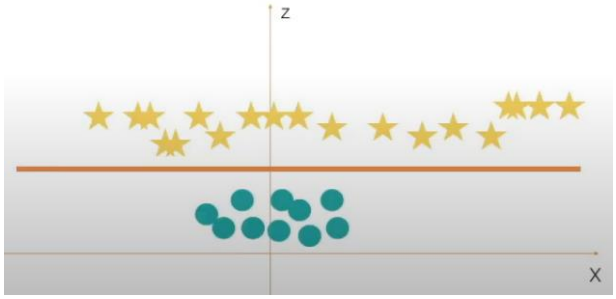


Figura 9. Ejemplo de SVM no lineal parte 2.

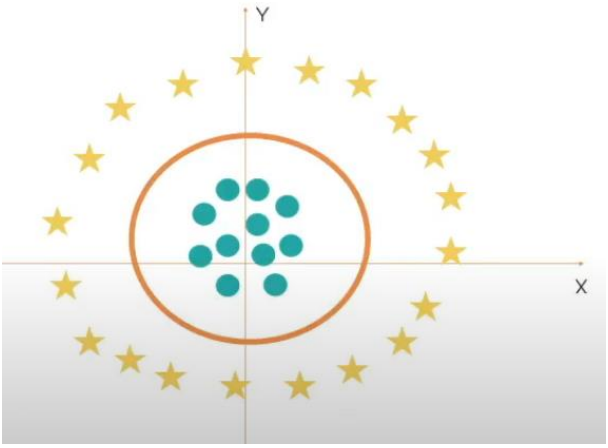


Figura 10. Ejemplo de SVM no lineal parte 3.

En el caso de que no sea un problema de clasificación lineal puede utilizar las funciones de kernel no lineales estas pasan los datos a un espacio (x,z) luego se transforma al espacio (x,y) [23]

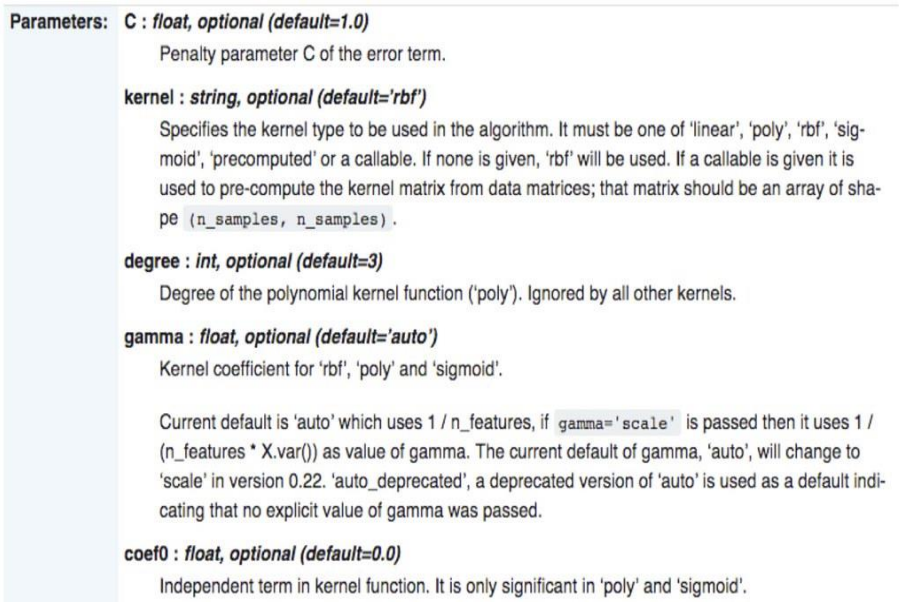
De manera similar al anterior algoritmo la librería Scikit Learn de python proporciona un modelo para la implementación SVM de clasificación,

en primer lugar, debemos de definir el módulo en este caso es sklearn.svm y la clase ubicada dentro del SVC también puede ser usados para regresión, el código es un Ejemplo de SVM [23]

```
from sklearn.svm import SVC
x_entrenamiento = variablesIndependientes_entrenamiento
y_entrenamiento = variableDependiente_entrenamiento
x_prueba = variablesIndependientes_prueba
y_prueba = variableDependiente_prueba

algoritmo = SVC()
algoritmo.fit(x_entrenamiento, y_entrenamiento)
algoritmo.predict(x_prueba)
```

Se pasa a definir las variables x e y para implementar los métodos de entrenamiento y prueba asi mismo posee parámetros por defecto que pueden ir cambiando para adecuar nuestro modelo como el Kernel que contiene funciones útiles para problemas no lineales o el de penalización que representa la clasificación errónea indica a la optimización del modelo cuanto error es soportable [23] El resto de parámetros pueden ser modificados acorde con las decisiones que tome el desarrollador para mejorar el modelo.



Parameters:

C : *float, optional (default=1.0)*

Penalty parameter C of the error term.

kernel : *string, optional (default='rbf')*

Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to pre-compute the kernel matrix from data matrices; that matrix should be an array of shape `(n_samples, n_samples)`.

degree : *int, optional (default=3)*

Degree of the polynomial kernel function ('poly'). Ignored by all other kernels.

gamma : *float, optional (default='auto')*

Kernel coefficient for 'rbf', 'poly' and 'sigmoid'.

Current default is 'auto' which uses 1 / n_features, if `gamma='scale'` is passed then it uses 1 / (n_features * X.var()) as value of gamma. The current default of gamma, 'auto', will change to 'scale' in version 0.22. 'auto_deprecated', a deprecated version of 'auto' is used as a default indicating that no explicit value of gamma was passed.

coef0 : *float, optional (default=0.0)*

Independent term in kernel function. It is only significant in 'poly' and 'sigmoid'.

Figura 11. Parámetros de SVM.

- **Agrupación de k-medias:** Pertenece al tipo de aprendizaje no supervisado caracterizado por su simpleza y capacidad de calcificación basado en la teoría de bayes, esta considera las características partículas de una clase es independiente del resto, consideremos la solicitud de un préstamo bancario como deseable o no dependiendo de su sueldo, historial crediticio, transiciones realizadas, entre otras características son consideras de manera independiente [23]

  Otro ejemplo, las votaciones de un país en la cual suponemos que el candidato A gano con 50.2% y el candidato B perdió con 49.8% en diferencias tan reducidas si un porcentaje de los electores hubiera elegido diferente hubieran obtenido otros resultados, de ese pequeño número de cómo podemos encontrar un grupo de personas? ¿Como las atraemos hacia un determinado grupo? La respuesta a ello es la agrupación como primer paso es recopilar información de cualquier tipo de los votantes que nos de algún indicio de que es relevante para ellos y sus influencias al votar, todo esto se coloca en el algoritmo de agrupación y cuando se entre a la campaña política se evaluara para saber si funciona nuestro modelo [21]

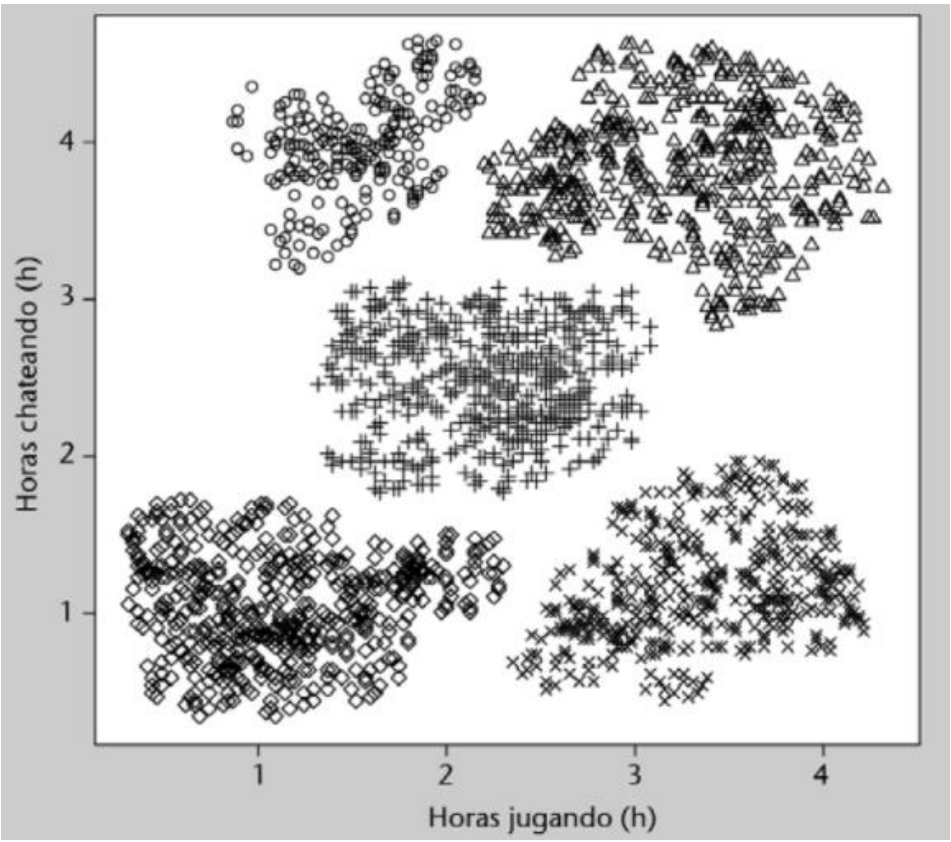


Figura 12. Actividad de jugadores online.

La imagen es una representación en la que se muestra 5 categorías o clases con diferencias notables marcado con símbolos distintos, usualmente en la realidad no de da un alejamiento de cada uno notorio, pero sirve en la ejemplificación de este algoritmo [21]

- **Redes Neuronales Artificiales ANN:** Las ANN son las que ofrecen la tarea de clasificación en conjuntos de datos debidamente etiquetados y regresión mediante datos continuos, así mismo de segmentación en datos sin etiquetar [24]

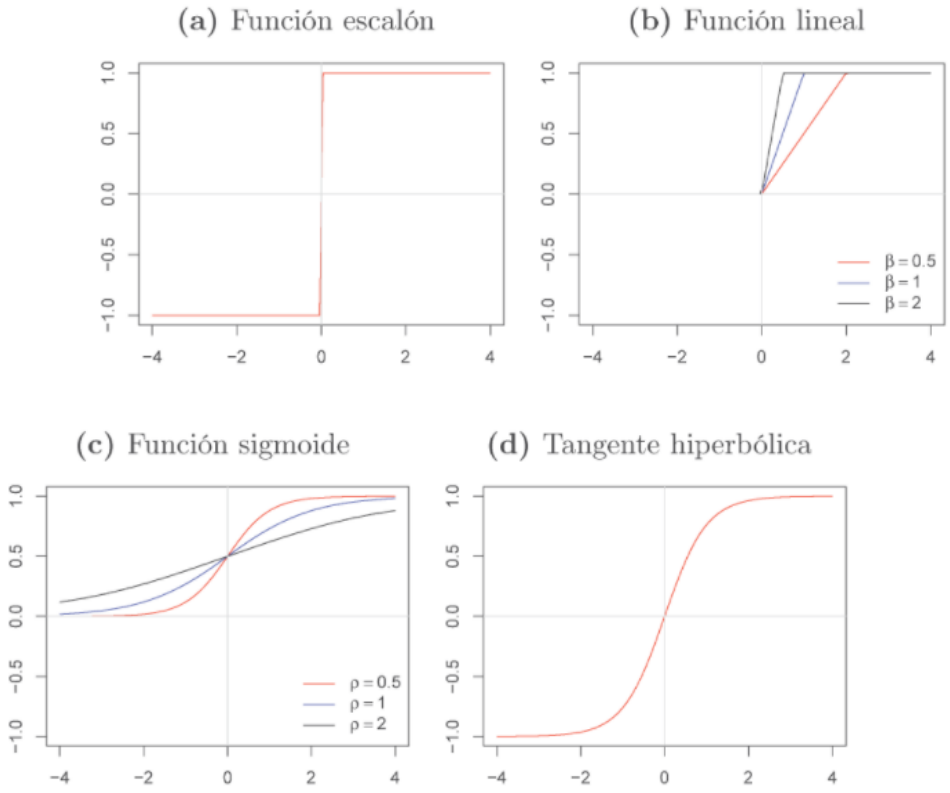


Figura 13. Representación de las funciones escalón, lineal, sigmoide e hiperbólica.

El aprendizaje profundo puede resolver tanto los problemas que aparecen en el caso A y B en la primera clasificación de imágenes se cuestionara si existe un coche y en el segundo si detecta objetos al frente o al costado [24]

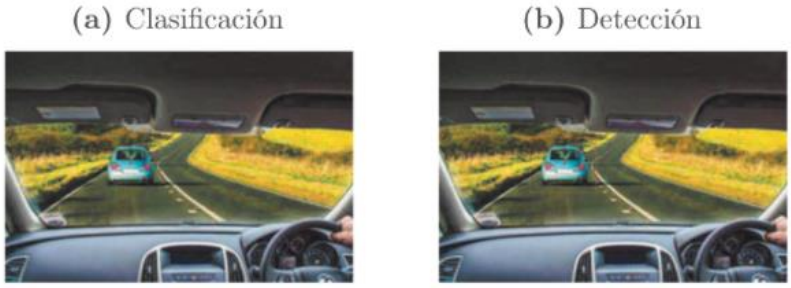


Figura 14. Resolver problemas A y B mediante aprendizaje profundo.

1.5.5. Cáncer de pulmón

El cáncer es una enfermedad que puede afectar a cualquier parte del cuerpo humano, está compuesto por células cancerosas, las células cancerosas exhiben una mitosis acelerada y descontrolada, lo que lleva a la aparición de una gran cantidad de tumores en el organismo huésped y aumenta su tamaño [3]

Este ejemplar del cáncer tiene origen en los pulmones, donde se generar un crecimiento anormal de células, estas aparecen generando nódulos en el órgano, los cuales en términos simplificados hace alusión a lesiones redondeadas, circunscritas y profundas, estos pueden ser benignos e indoloros, como malignos, de igual modo representan un riesgo para el correcto funcionamiento del pulmón, existen dos tipos de cáncer al pulmón [3]

Cáncer de pulmón no microcítico

[3] Gran cantidad de los canceres de pulmón son no microcítico, entre el 80% y 85%, debido a las diferentes células pulmonares en las que se originan existen subtipos, esta se encuentra bajo a la misma categoría debido a que tanto el pronóstico como el tratamiento son similares:

- Adenocarcinoma
- Carcinoma de las células escamosas
- Carcinoma de las células grandes
- Otros subtipos

Cáncer de pulmón microcítico

Esta categoría de cáncer pulmonar represente la minoría restantes de los identificados, que son entre 10% a 15%, es conocida como cáncer de células en avena, su propagación es muy rápida en comparación con el cáncer no microcítico, por ello es difícil de erradicar, aunque es tratable con quimioterapia y radioterapia, suele tener rebrotes en su huésped [3]

[3] Este tipo de enfermedad involucran cirugía, medicinas y radiación; estas pueden resultar en la extracción completa del miembro afectado o parte de esta, no siempre se genera un tumor el extraer, como en caso del cáncer de sangre el cual es solo tratado con medicamentos, estas son suministrados vía oral o intravenosa acorde con del tipo de cáncer a tratar, los medicamentos empleados particularmente en la detección son:

- Quimioterapia
- Terapia dirigida
- Inmunoterapia
- Terapia hormonal

Estas disminuyen el crecimiento de las células y la también las eliminan, por lo general traen efectos secundarios como la pérdida de cabello, afectar directamente en el estado de salud del paciente [3]

Detección del cáncer:

Usualmente se realizan estudios con imágenes medidas para encontrarlo, estos hacen uso de rayos x, sustancias radiactivas, campos magnéticos y ondas sonoras con el fin de generar una imagen de la parte interna del cuerpo, los pacientes se realizan estas pruebas para el diagnóstico de posibles áreas sospechosas, conocer la propagación del cáncer, ayudar a seleccionar un procedimiento adecuado y finalmente si posibles signos de este tipo regresan después del tratamiento [3] Entre los más usados a través de imágenes médicas son:

- **Tomografía por emisión de positrones (PET):** En este estudio se introduce en la sangre un tipo de azúcar ligeramente radiactivo que se acumula en las células cancerosas, también se puede combinar con

las TC mediante el uso de una maquina especial que otorga al médico comparar las zonas de mayor radiactividad de PET con las imágenes más detalladas de TC [3]

- **Tomografía computarizada TC:** La probabilidad de diagnóstico es mayor a las técnicas convencionales hacen uso de los rayos x producen img transversales del cuerpo a diferencia de una radiografía esta toma varias imágenes y luego el ordenador las combina para mostrar tamaño, forma, posición, una sección de la parte del cuerpo y buscar si se ha propagado [3]
- **Gammagrafía ósea:** Este estudio identifica si el cáncer se ha propagado hasta los huesos inyectando un reducido porcentaje de radiactividad en la sangre que se acumula en las áreas anormales de los huesos, esto no es necesario ya que las PET también pueden mostrar esto [3]
- **Imágenes por resonancia magnética MRI:** De manera similar a las TC estas ofrecen imágenes detalladas de los tejidos blandos del cuerpo utilizando ondas de radio e imanes potentes en lugar de rayos X, se usa especialmente para diagnosticar la programación del cáncer a la medula espinal o cerebro [3]
- **Radiografía de pecho:** Esta es una de las primeras que generalmente se les solicita para identificar zonas anormales en el pulmón, en caso no logre identificar algo o tenga ciertas dudas los profesionales de la salud le indicaran que realice otras pruebas [3]

1.5.6. Herramientas de implementación:

En el desarrollo e implementación se deberá seleccionar el lenguaje, librerías y entorno de desarrollo (IDE), existen variedad de opciones y generalmente con los requisitos mencionados ya que se proporciona la implementación de los algoritmos de ML, algunas recomendaciones a tener en cuenta son usar el idioma de su dominio entre los admitidos del aprendizaje automático, la selección del IDE ya que dependerá de la familiaridad y su nivel de comodidad con este, respeto a las plataformas la gran parte de ellas son libres no obstante en ocasiones puedes requerir alguna licencia para adquirir mayor cantidad de uso (Benítez, 2014).

Los lenguajes de programación: En este apartado se muestra una lista de

los idiomas admitidos en ML

- C
- C ++
- Julia
- Matlab
- Octava
- Pitón
- R

Es relevante aclara que existen otros por lo que no está completa, pero esta lista cubre los lenguajes más populares usados en el desarrollo e implantación del aprendizaje automático acorde a ello con las recomendaciones planteadas anteriormente se deberá seleccionar uno [21]

El entorno de desarrollo (IDE): Se contemplan a aquellas que pueden implementar las aplicaciones de ML, de manera similar al anterior punto no está completa la lista y se anima a probar los entornos ante se elegir solo uno [21]

- Anaconda
- Cuaderno iPython / Jupyter
- Google –Colab
- Julia
- Pycharm
- R Studio
- Rodeo
- Spyder

Plataformas: La lista que se muestra a continuación no es absoluta por lo que es recomendable registrarse en cada servicio, probar y seleccionar alguno [21] Aquellas plataformas donde se pueden implementar las aplicaciones de Ml son las siguientes:

- IBM
- Microsoft Azure

- Google Cloud
- Amazonas
- Mlflow

II. MATERIALES Y MÉTODO

2.1. Tipo y Diseño de Investigación

La investigación es cuantitativa debido a que se emplearan métodos matemáticos y estadísticos para indicar y medir los resultados de la hipótesis planteada [25] Es decir, se trabaja con valores numéricos para evaluar el consumo de recursos y rendimiento. Se considera cuasi experimental el diseño de la investigación ya que aún no se realizará la etapa de experimentos o pruebas no obstante la autenticidad de los datos presentados no debilita la investigación [25] Además, la investigación es tecnológica y aplicada, la primera es considerada por que se afrontan los problemas de carácter tecnológico, se obtendrá un producto tecnológico y se utilizan investigaciones con conocimiento teórico científico y aplicada ya que a su vez se emplearán conocimientos teóricos probados al problema planteado [26]

2.2. Variables, Operacionalización

Población: La población está determinada por los modelos de aprendizaje profundos más sobresalientes empleados para la clasificación del conjunto de datos ImageNet.

Muestra: Para determinar la muestra será a conveniencia por lo que se selecciona las arquitecturas Mobilenetv2, Densenet201, Efficientb4 y Resnet50 por ser los más sobresalientes en las métricas de evaluación del conjunto de datos ImageNet y para la clasificación del cáncer de pulmón en la investigación.

Tabla 1. Indicadores de la variable dependiente e independiente.

Variables de estudio	**Definición Conceptual**	**Definición operacional**	**Dimensiones**	**Indicadores**	**Ítem**	**Instrumento**	**Valores finales**	**Tipo de variable**	**Escala de medición**
Variable independiente Algoritmos de aprendizaje automático	Los algoritmos de aprendizaje automático son una categoría de algoritmos informáticos que utilizan técnicas estadísticas y de análisis de datos para identificar	Seleccionar una muestra de datos que incluya ejemplos de la tarea específica que se desea mejorar con el algoritmo de aprendizaje automático. Seleccionar uno o más algoritmos de aprendizaje automático que sean adecuados	Consumo de recursos	Grado de consumo de GPU	$ce = \sum_j^n \frac{ce_j}{n}$	Ficha digital de observación	Gigabyte (GB)	Numérica	Razón (0GB-10GB)
				Grado de consumo de memoria	$cm = \sum_j^n \frac{cm_j}{n}$				
				Promedio de tiempo de respuesta	$Tr = \sum_j^{n_f} \frac{tf_j - tf_i}{n}$		Minutos (Min)		Proporcional (0%-100%)

	patrones en los datos de entrada, con el objetivo de aprender y mejorar su desempeño en la tarea específica para la que fueron diseñados, sin ser explícitamente programados para ello	para la tarea específica. Entrenar el algoritmo utilizando los datos de la muestra, ajustando los parámetros del algoritmo para maximizar su desempeño en la tarea. Evaluar el desempeño del algoritmo utilizando una métrica adecuada para la tarea específica, como							

		precisión, recall, F1-score, etc. Repetir los pasos 3 y 4 varias veces utilizando diferentes conjuntos de datos de entrenamiento y validación, con el objetivo de obtener una medida más robusta del desempeño del algoritmo. Analizar los resultados obtenidos y comparar el desempeño de diferentes							

		algoritmos de aprendizaje automático para la misma tarea, si se están evaluando varios algoritmos.							
Variable dependiente Clasificación de cáncer de pulmón	Se refiere a la evaluación y categorización de los tumores pulmonares según su tipo y grado de malignidad, utilizando criterios médicos y patológicos establecidos por la	Se puede medir mediante un conjunto de procedimientos médicos y patológicos utilizados para evaluar y categorizar los tumores pulmonares según su tipo y grado de malignidad. Estos procedimientos pueden incluir pruebas de imagen, biopsias,	Rendimiento	Exactitud	$E = \frac{VP + VN}{VP + VN + FP -}$		Porcentaje (%)	Numérica	Proporcional (0%-100%)
				Precisión	$P = \frac{VP}{VP + FP}$				
				Recall	$R = \frac{VP}{VP + FN}$				
				F	$F = 2 \cdot \frac{P \cdot E}{P + E}$				

	comunidad científica. Esta definición se basa en una revisión de la literatura médica y oncológica relacionada con el diagnóstico y tratamiento del cáncer de pulmón.	análisis histológicos y moleculares, y la aplicación de clasificaciones estandarizadas, como la clasificación TNM (Tumor, Nodes, Metastasis). La variable puede medirse de manera cualitativa o cuantitativa, dependiendo del enfoque de la investigación y los objetivos de la medición.							

2.3. Población de estudio, muestra, muestreo y criterios de selección

Población: La población está determinada por los modelos de aprendizaje profundos más sobresalientes empleados para la clasificación del conjunto de datos ImageNet.

Muestra: Para determinar la muestra será a conveniencia por lo que se selecciona las arquitecturas Mobilenetv2, Densenet201, Efficientb4 y Resnet50 por ser los más sobresalientes en las métricas de evaluación del conjunto de datos ImageNet y para la clasificación del cáncer de pulmón en la investigación.

2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad

Ficha digital de observación: Este informe estará automatizado mediante métodos que estarán incorporados dentro del código fuente que se compilara y a partir de eso se genera un matriz de confusión en caso se requiera (ver anexo n°2) o también podemos recurrir a la invocación de los métodos de los indicadores de rendimiento (ver anexo n°3) los cuales luego permitirán hacer los cálculos de cada uno de ellos al mismo tiempo del indicador de tiempo de respuesta, en el caso de los indicadores de consumo de recursos (ver anexo n°4) restantes se apoyara mediante el reporte del administrador de tareas.

2.5. Procedimiento de análisis de datos

2.5.1. Consumo de recursos.

Grado de consumo de GPU: Este indicador permitirá medir el grado de rendimiento del ordenador al momento de ejecutarse la etapa de experimentación, la fórmula se expresa de la siguiente manera:

$$ce = \sum_{j}^{n} \frac{ce_j}{n}$$

Donde:

ce: Grado de consumo de GPU

ce_j: Grado de consumo de GPU en la prueba j

n: Es el total de pruebas

Grado de consumo de memoria: El segundo indicador nos mostrar la memoria consumida durante las pruebas del modelo, la fórmula se expresa de la siguiente manera:

$$cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Donde:

cm: Grado de consumo de memoria

cm_j: Grado de consumo de memoria en la prueba j

n: Es el total de pruebas.

Promedio de tiempo de respuesta: De manera similar como indica su nombre este indicador es el tiempo promedio que está en ejecución el modelo en la etapa de pruebas, la fórmula se expresa de la siguiente manera:

$$Tr = \sum_{j}^{n_f} \frac{tf_j - tf_i}{n}$$

Donde:

Tr:Tiempo de respuesta

tf_j:Tiempo de respuesta final

tf_i:Tiempo de respuesta inicial

n: Es el total de pruebas

2.5.2. Consumo de rendimiento:

Exactitud: Esta es la parte de la predicción que el modelo hace correctamente en el caso de obtener un porcentaje del 91% a primera impresión esto indicaría que es un buen resultado no obstante el modelo solo identifica 1 de 9 objetos a clasificar [27] La fórmula se expresa de la siguiente manera, además los valores VP, VN, FP, FN se obtendrán de la invocación del método de matriz de confusión o invocar directamente a la exactitud.

$$E = \frac{VP + VN}{VP + VN + FP + FN}$$

E: Exatitud

Donde:

VP: Verdadero Positivo

VN: Verdadero Negativo

FP: Falso Positivo

FN: Falso Positivo

Precisión: Según [27] es la proporción de identificaciones positivas correcta, suponiendo que obtenemos un valor de 0.5 significa que el modelo acierta el 50% de las veces, la formula se expresa de la siguiente manera e igualmente que el indicador anterior podemos invocar al método.

$$P = \frac{VP}{VP + FP}$$

Donde:

P: Precisión

VP: Verdadero Positivo

FP: Falso Positivo

Recall: Es también llamado sensibilidad o exhaustividad responde al porcentaje de positivos reales que se idéntica correctamente [27] Usualmente existe un problema entre precisión y exhaustividad, al mejorar una se reduce la otra y viceversa, la formula se enuncia de la siguiente forma y sus valores se obtendrán de la invocación de su respectivo método.

$$R = \frac{VP}{VP + FN}$$

Donde:

R: Recall

VP: Verdadero Positivo

FN: Falso Positivo

F: Es llamada F-score o puntuación F, es una medida de presión de una prueba de la precisión y recall aplicable cuando los valores son tasas [27] Alcana su mejor valor en 1 eso indica tanto que P y E son perfectos, caso contrario se obtendrá cero, la formula puede ser invocada y se expresa como:

$$F = 2 \cdot \frac{P \cdot E}{P + E}$$

Donde:

F: F-score

P: Precisión

E: Exactitud

2.6. Criterios éticos

2.6.1. Confiabilidad

Los datos y autoría usada en esta investigación deben de adquirirse de manera licita evitando perjudicar al titular [28]

2.6.2. Conformabilidad

Acorde con “El Código Deontológico Del Colegio De Ingenieros De Perú” 2012 en su tercer capítulo “En temas de ingeniería los ingenieros serán veraces y objetivos en sus afirmaciones y escritos competitivos” la cual se contemplará [29].

2.6.3. Integridad

En la investigación se utilizó el principio ético de honestidad e integridad, puesto que no se manipularán los datos a conveniencia, mostrando resultados objetivos asimismo se referencio las fuentes de información para no incidir en el plagio.

III. RESULTADOS Y DISCUSIÓN

3.1. Resultados

Se muestran los resultados de consumo de recursos computacionales durante la ejecución de los modelos establecidos en relación al tiempo de ejecución, consumo de GPU y memoria por arquitectura.

Tabla 2. Tiempo de respuesta y grado de consumo.

Descripción	Arquitectura	Promedio de tiempo de respuesta	Grado de consumo de GPU (GB)	Grado de consumo de memoria (GB)
Con Aumento de datos	resnet50_DA	35min	8.86	2.75
	efficientb4_DA	1h 4min 16s	8.86	2.87
	densenet201_DA	34min 3s	8.86	3.17
	mobilenetv2_DA	57min 7s	8.86	3.61
Sin Aumento de datos	resnet50	18min 8s	8.86	2.82
	efficientb4	26min 9s	8.86	3.30
	densenet201	21min 31s	8.86	4.42
	mobilenetv2	18min	8.89	2.66

Nota: El consumo de recursos computacionales fue calculado por medio de los recursos del entorno de trabajo.

Las métricas de evaluación establecidas para el entrenamiento, validación y pruebas se muestran para cada arquitectura y la descripción indica si se usó el aumento de datos.

Tabla 3. Evaluación de métricas en el entrenamiento y validación.

Descripción	Entrenamiento y validación	Acurracy	Precision	Recall	F1 Score
Con Aumento de datos	resnet50_DA	96.17%	93.03%	91.52%	90.90%
	efficientb4_DA	98.86%	98.35%	97.06%	95.40%
	densenet201_DA	99.76%	99.67%	99.35%	99.05%
	mobilenetv2_DA	99.47%	99.34%	98.53%	98.98%
Sin Aumento de datos	resnet50	99.43%	98.86%	98.86%	98.06%
	efficientb4	99.51%	99.34%	98.69%	97.62%
	densenet201	99.84%	99.67%	99.67%	98.84%
	mobilenetv2	99.88%	99.84%	98.67%	99.76%

Nota: Las métricas de evaluación fueron calculados llamando a la función de cada una por medio del entorno de trabajo. Fuente: Elaboración propia

Tabla 4. Evaluación de métricas en pruebas.

Descripción	Pruebas	Acurracy	Precision	Recall	F1score
Con Aumento de datos	resnet50_DA	95.08%	91.48%	88.57%	89.99%
	efficientb4_DA	95.32%	91.29%	89.84%	90.54%
	densenet201_DA	95.00%	92.00%	87.62%	89.79%
	mobilenetv2_DA	91.43%	85.08%	79.68%	82.38%
Sin Aumento de datos	resnet50	91.43%	83.28%	82.22%	82.71%
	efficientb4	93.02%	87.21%	84.44%	85.77%
	densenet201	88.10%	77.97%	73.02%	75.34%
	mobilenetv2	54.92%	0.07%	0.06%	0.07%

Nota: Las métricas de evaluación fueron calculados llamando a la función de cada una por medio del entorno de trabajo. Fuente: Elaboración propia

En las configuraciones previas realizadas antes de entrenar y validar los modelos se indican las métricas, tasa de entrenamiento, el optimizador, función de perdida, parada temprana, pasos por época, épocas (limite) y épocas ejecutadas, en retrollamada o callbacks se le agrega el punto de control (Check Pointer) y parada o detención temprana (Early Stopping)

Tabla 5. Configuraciones antes del entrenamiento.

Arquitectura	Métricas	Tasa de entrenamiento	Optimizador	Función de perdida	Parada temprana
resnet50_DA efficientb4_DA densenet201_DA mobilenetv2_DA resnet50 efficientb4 densenet201 mobilenetv2	Accuracy, Precision, Recall, F1_score	0.00001	Adam	Categorical crossentropy	15

Nota: Las configuraciones establecidas son definidas antes de entrenar y validar el modelo.

Fuente: Elaboración propia

Tabla 6. Configuraciones antes del entrenamiento y número de ejecuciones

Arquitectura	Pasos por Épocas	Épocas	Retrollamada	Épocas Ejecutadas
resnet50_DA efficientb4_DA	20	200	tb_resnet50_DA tb_efficientb4_DA	68 112

densenet201_DA	tb_densenet_DA	60
mobilenetv2_DA	tb_mobilenetv2_DA	115
resnet50	tb_resnet50	82
efficientb4	tb_efficientb4	92
densenet201	tb_densenet	76
mobilenetv2	tb_mobilenetv2	157

Nota: Las configuraciones establecidas son definidas antes de entrenar y validar el modelo, adicionalmente se agregan el número de épocas una vez culminada la ejecución. Fuente: Elaboración propia

En la figura N° 15 se muestra los resultados de todos los modelos en el entrenamiento y validación definidos con el indicador de Exactitud en la que se aprecia solo 6 arquitecturas densenet201_DA, mobilenetv2_DA, resnet50, efficientb4, densenet201, mobilenetv2 alcanzan una exactitud del 99% en clasificar los tipos de cáncer de pulmón.

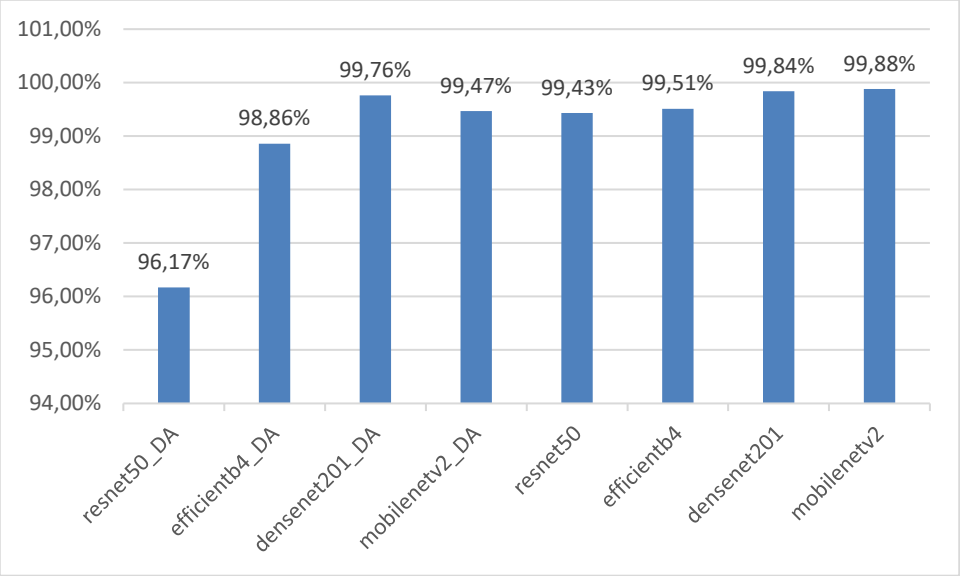


Figura 15. Entrenamiento y validación: Acurracy.

En la figura N° 16 se visualiza los resultados de todos los modelos definidos en el entrenamiento y validación con el indicador de Precision en la que se aprecia solo 5 arquitecturas densenet201_DA, mobilenetv2_DA, efficientb4, densenet201, mobilenetv2 alcanzan una precision del 99% en clasificar los tipos de cáncer de pulmón.

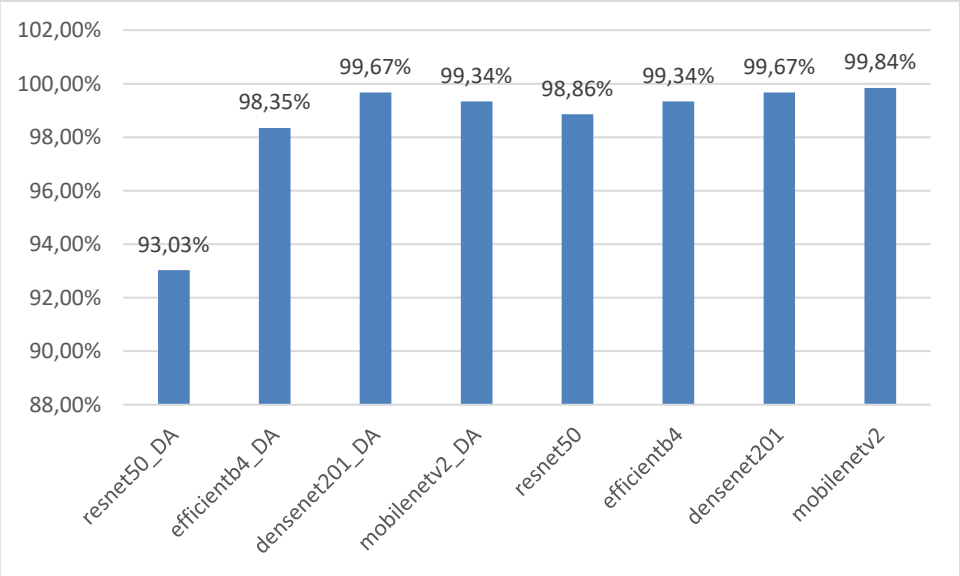


Figura 16. Entrenamiento y validación: Precision.

En la figura N° 17 se visualiza los resultados de todos los modelos definidos en el entrenamiento y validación con el indicador de Recall en la que se aprecia solo 4 arquitecturas densenet201_DA, mobilenetv2_DA, densenet201 y mobilenetv2 alcanzan una sensibilidad del 99% en clasificar los tipos de cáncer de pulmón.

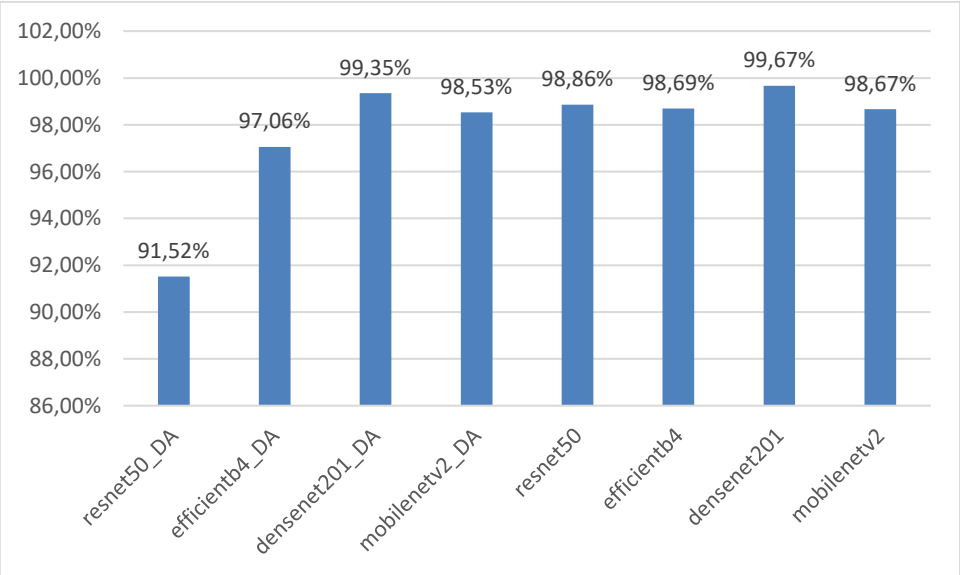


Figura 17. Entrenamiento y validación: Recall.

En la figura N° 18 se visualiza los resultados de todos los modelos definidos en el entrenamiento y validación con el indicador de F1 Score en la que se aprecia solo 2 arquitecturas densenet201_DA y mobilenetv2 alcanzan el valor F1 del 99% en clasificar los

tipos de cáncer de pulmón.

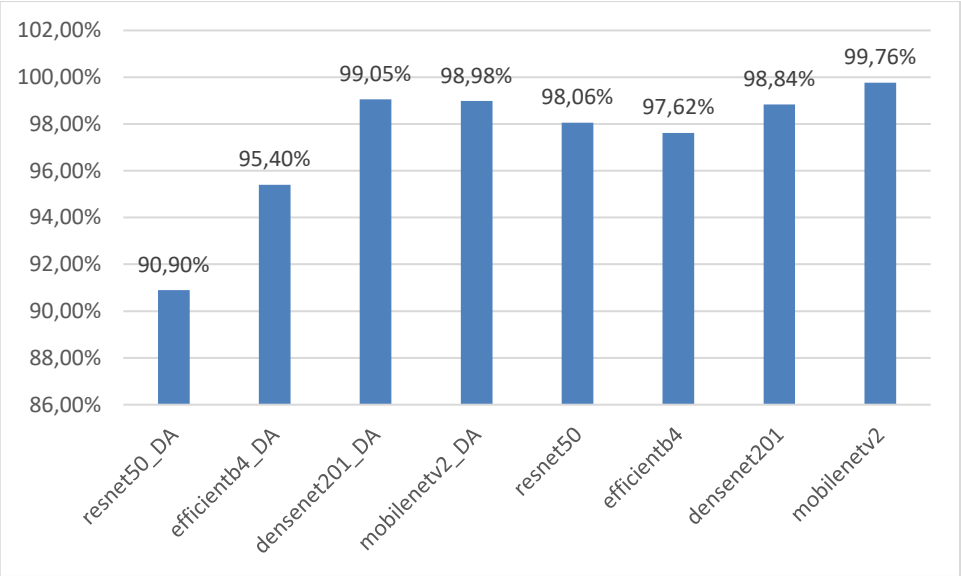


Figura 18. Entrenamiento y validación: F1 Score.

En la figura N° 19 se visualiza los resultados de todos los modelos en las pruebas definidos con el indicador de Exactitud en la que se aprecia solo 3 arquitecturas resnet50_DA, efficientb4_DA y densenet201_DA alcanzan una exactitud del 95% en clasificar los tipos de cáncer de pulmón.

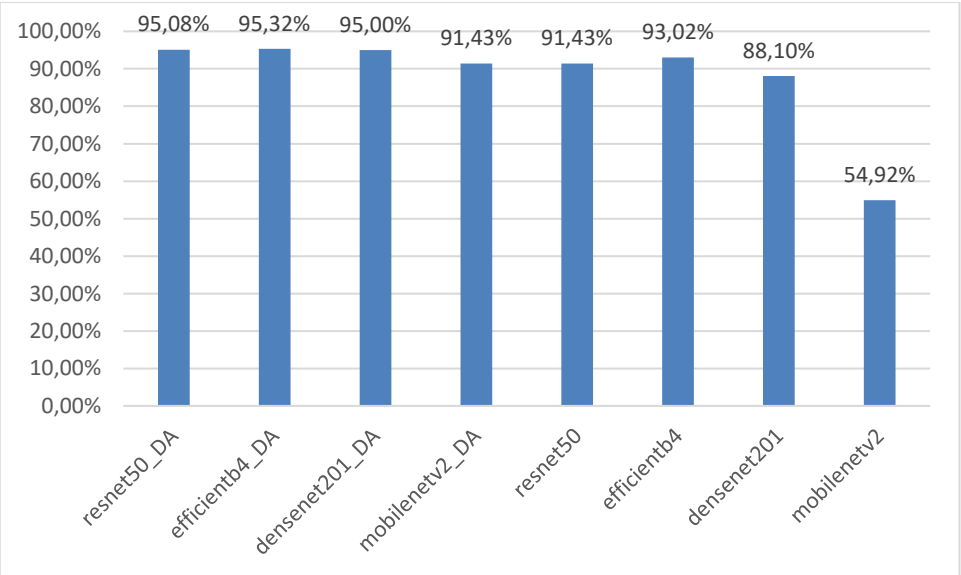


Figura 19. Pruebas: Acurracy.

En la figura N° 20 se visualiza los resultados de todos los modelos en las pruebas definidos con el indicador de Precision en la que se aprecia solo 1 arquitectura densenet201_DA

alcanzan una precisión del 92% y 91% resnet50_DA y efficientb4_DA en clasificar los tipos de cáncer de pulmón

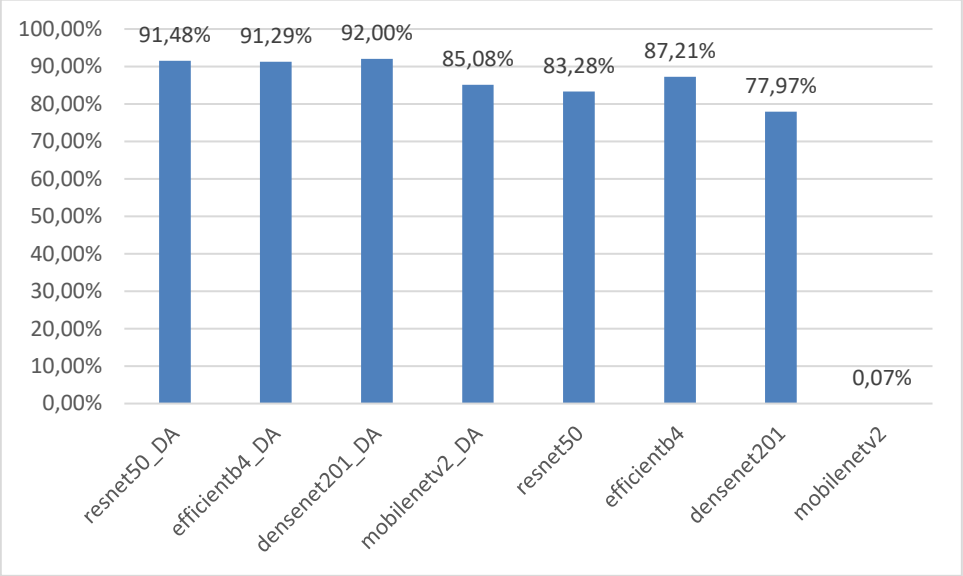


Figura 20. Pruebas: Precision.

En la figura N° 21 se visualiza los resultados de todos los modelos en las pruebas definidos con el indicador de Recall en la que se aprecia solo 1 arquitectura efficientb4_DA alcanzan una sensibilidad del 89% en clasificar los tipos de cáncer de pulmón.

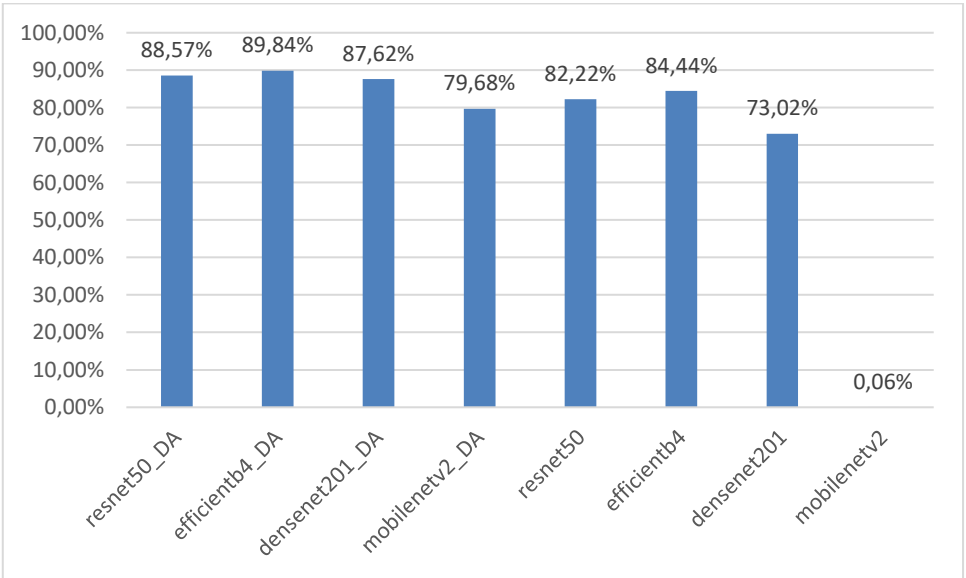


Figura 21. Pruebas: Recall.

En la figura N° 22 se visualiza los resultados de todos los modelos en las pruebas definidos con el indicador de F1 Score en la que se aprecia solo 1 arquitectura efficientb4_DA alcanzan un valor f del 90% en clasificar los tipos de cáncer de pulmón

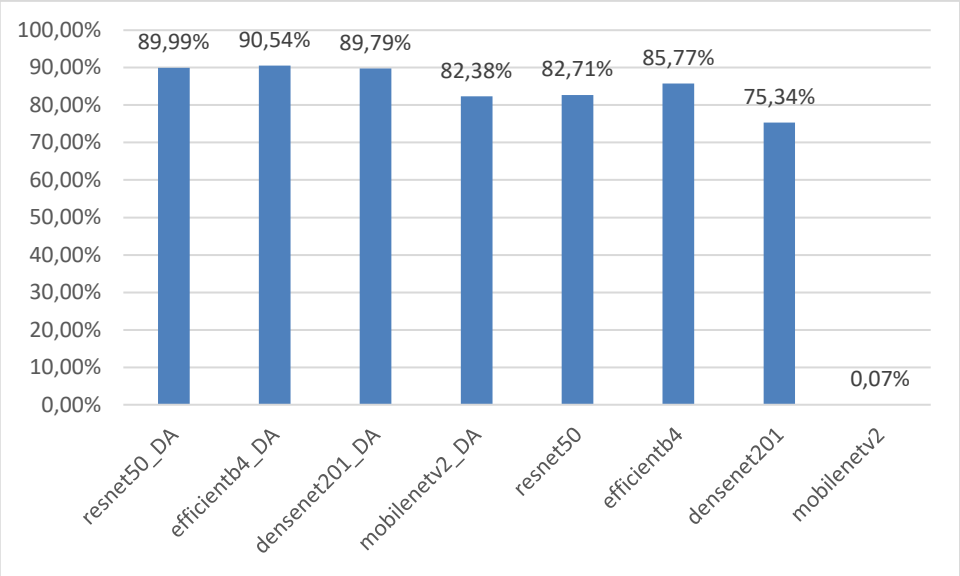


Figura 22. Pruebas: F1 Score.

En la figura N° 23 se aprecia los resultados de todas las métricas del entrenamiento y validación con aumento de datos para clasificar los tipos de cáncer de pulmón

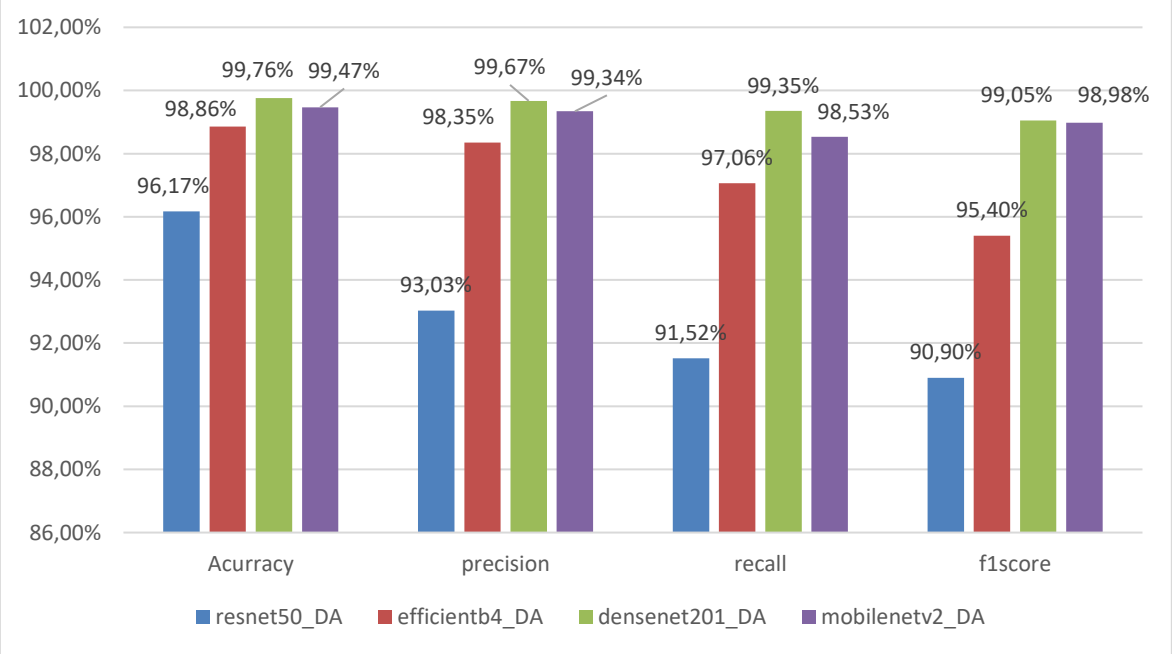


Figura 23. Métricas de evaluación en entrenamiento y validación con aumento de datos.

En la figura N° 24 se visualiza los resultados de todas las métricas de pruebas con aumento de datos para clasificar los tipos de cáncer de pulmón

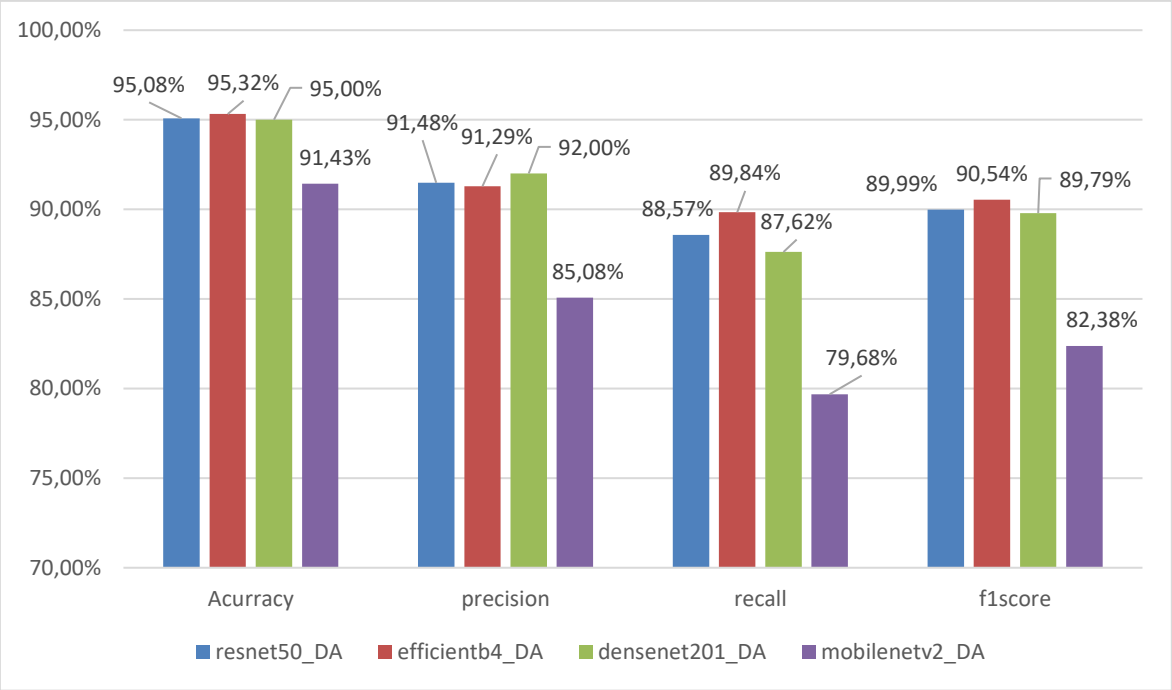


Figura 24. Métricas de evaluación en pruebas con aumento de datos.

En la figura N° 25 se visualiza los resultados de todas las métricas del entrenamiento y validación sin aumento de datos para clasificar los tipos de cáncer de pulmón

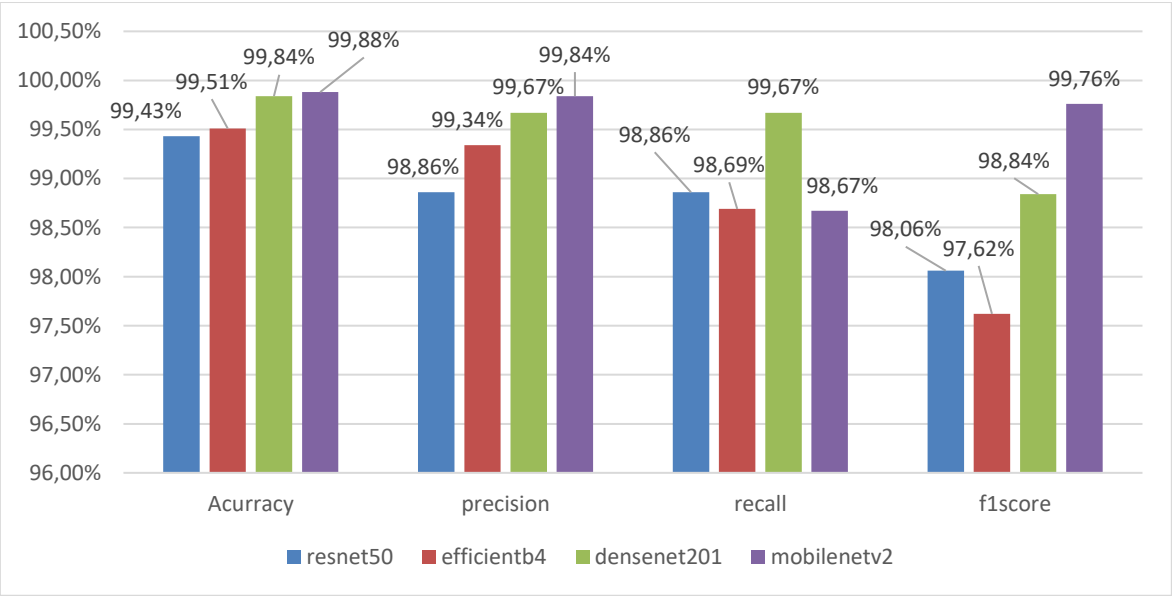


Figura 25. Métricas de evaluación en entrenamiento y validación sin aumento de datos.

En la figura N° 26 se visualiza los resultados de todas las métricas de pruebas sin aumento de datos para clasificar los tipos de cáncer de pulmón

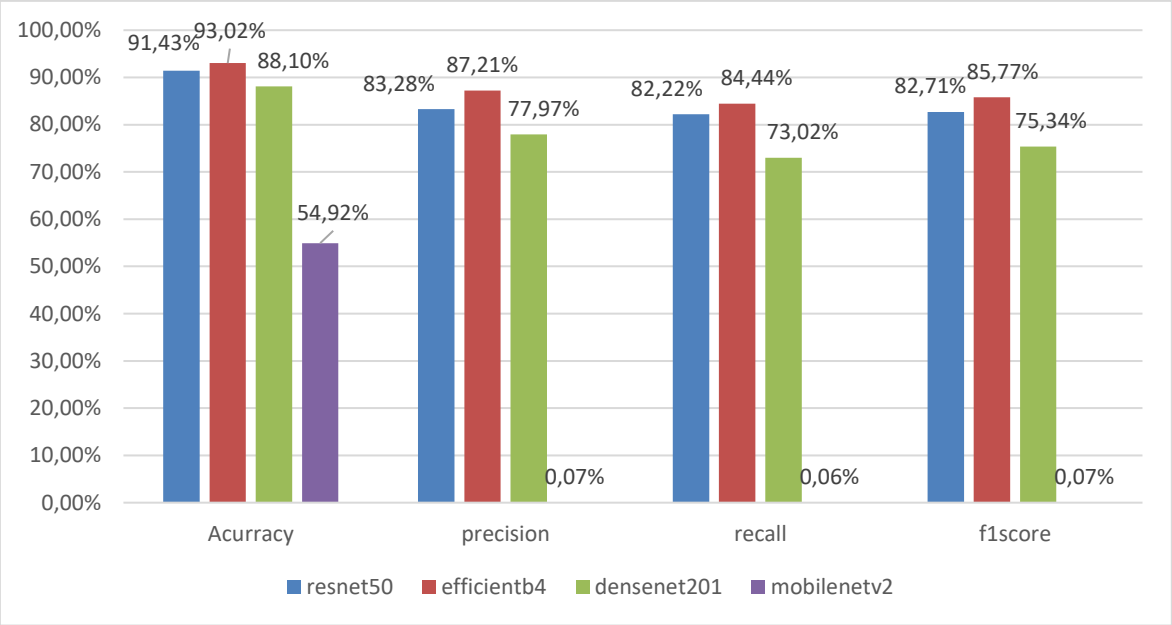


Figura 26. Métricas de evaluación en pruebas sin aumento de datos.

3.2. Discusión

Acorde con la investigación [4] usaron el conjunto de datos “LIDC” en su primera etapa consideran el preprocesamiento en la cual se aplica un filtro mediano y gaussiano el primero para eliminar el ruido y no detectar falsos nódulos, después el segundo filtro suaviza y quita el ruido moteado o salpicado de la imagen, en la segunda etapa se realiza la segmentación para extraer la región de interés, luego en la tercera etapa extraen las características (área, perímetro, centroide, diámetro, excentricidad e intensidad) las cuales son empleadas en la etapa final de clasificación (Maligno/Benigno) usando como clasificador SVM obteniendo una exactitud del 92% en detección y 86% en clasificación. Mientras que en nuestra investigación se optó por el conjunto de datos “Chest CT Scan Images” con 1000 img separadas por tipo de cáncer (Adenocarcinoma, Carcinoma de células escamosa, Carcinoma de células grandes y Normales) para el entrenamiento, prueba y validación, después se implementan las arquitecturas Resnet50_DA, Efficientb4_DA, Densenet20_DA y Efficientb4 después del entrenamiento en las pruebas se obtiene una exactitud del 95.08%, 95.32%, 95.00% y 93.02% correspondiente a cada arquitectura.

Acorde con la investigación [30] usa el conjunto de LIDC y de Kaggle, la metodología propuesta, consta de dos partes en la primera se detecta los nódulos pulmonares después se entrena y prueba el modelo, primero en el preprocesamiento se eliminan los vasos y tejidos innecesarios, se ajusta la imagen luego en la segmentación se emplea el umbral Otsu, después en la detección de bordes se filtra para obtener nitidez y suavidad en los bordes, en la región de interés los nódulos mayores a 3cm son considerados malignos, en la extracción de características GLCM, posteriormente se recortan los nódulos obteniendo 6691 (Benigno 4012, Maligno 2679), se extraen las características mediante el método “La matriz de co-ocurrencia de nivel gris” (GLCM) y se usa una red neuronal profunda (DNN) para clasificar obtenido una exactitud de 91% y 88.21% en entrenamiento y pruebas correspondientemente.

De forma similar acorde con la investigación [31] primero se obtiene las imágenes de tomográfica computarizada pasando al preprocesamiento de las misma es este punto se incluye suavizado, realce, segmentación, apertura morfológica y selección de región de interés (ROI), como tercer paso se realiza la extracción de características utilizando un análisis de textura basado en GLCM y un enfoque paramétrico estadístico y luego se calcularon los valores de las características, finalmente en la etapa de clasificación se obtienen el mejor rendimiento en las métricas de evaluación con el clasificador SVM 78.95% en exactitud, 77% precisión, 83% sensibilidad y 79% f1score.

Asimismo en la investigación [12] se realizó una detección del cáncer de pulmón basado en imágenes médicas (Tomografías computarizadas) con el método de matriz de coocurrencia de nivel gris (GLMC) y máquina de vectores de soporte (SVM) el cual consta de 4 etapas principales inicialmente se realiza el preprocesamiento para mejorar la calidad de las imágenes, segundo se segmenta para separar el objeto deseado del resto, luego se extraen las características (área, contraste, energía, entropía y homogeneidad) y finalmente clasifica como normal, benigno o maligno el diagnóstico del modelo en exactitud es de 83%.

En la presente investigación se obtenido resultados más óptimos respecto a la investigación [30] [31] [12] se ha considerado solo uno conjunto de datos dividido en 70% para en entrenamiento, 20% de prueba y 10% de validación, después seleccionamos las arquitecturas aplicamos aumento de datos o no en los modelos, realizamos una configuración previa antes de entrenar y finalmente realizamos una clasificación sobre la imagen obteniendo una exactitud de 98.86%, precisión 98.35% sensibilidad 97.06% y puntuación F 95.40% en el entrenamiento y una exactitud del 95.32% precisión del 91.29% sensibilidad del 89.84% y puntuación F del 90.54% en las pruebas.

3.3. Aporte de la investigación

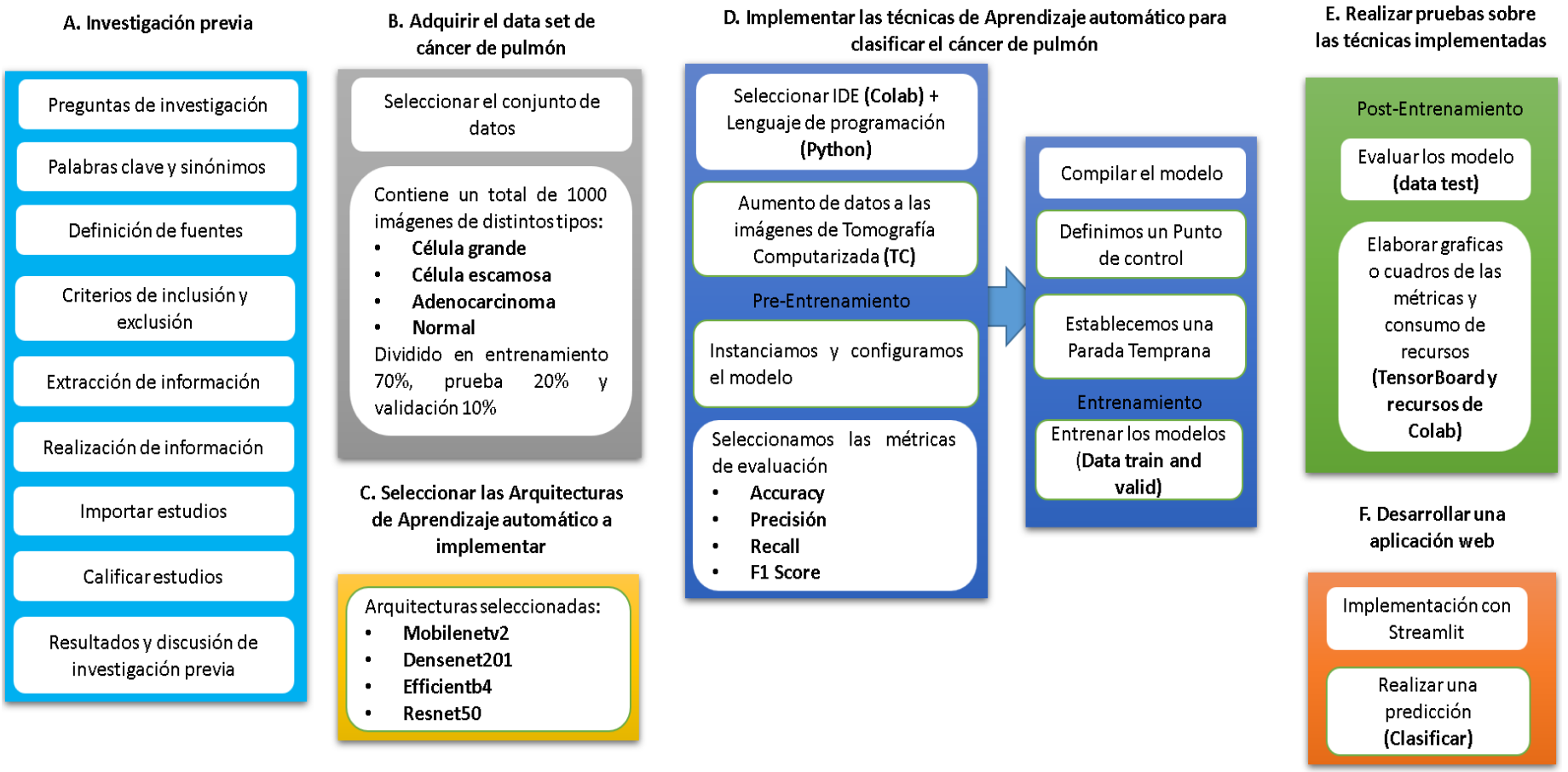


Figura 27. Metodología propuesta

3.3.1. Investigación Previa

Primero definimos una metodología de investigación de artículos, en el desarrollo del este trabajo de investigación se realizó mediante la automatización del software Parsifal costando de 4 partes revisión, planificación, conducción y reportes en la que primero se detalla aspectos generales pasando a los más abstractos, primeramente el título y su descripción, posteriormente en la planificación realizamos los objetivos, PICOC, preguntas de investigación, palabras clave y sinónimos, cadenas de búsqueda, fuentes y criterios de selección, en la etapa de verificación de evaluación de la calidad se definen las preguntas, respuestas y el puntaje de aceptación. En el formulario de extracción de datos consideramos autores, fuente, año, resumen, país y procedencia, en la conducción buscamos dentro de cada fuente y los resultados obtenidos son importados a su respectiva fuente y comienza el primer filtro de calificación de las investigaciones para luego calificarlas en base a las preguntas definidas y aquellas que pasen el puntaje especificados son consideradas en esta investigación en esta etapa se pueden extraer información y generación de graficas estadísticas, finalmente en reportes se generaran un archivos con la información que seleccionemos de nuestra investigación.

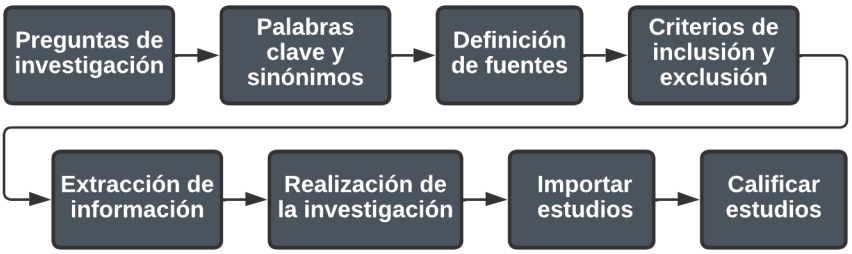


Figura 28. Proceso de revisión sistemática.

3.3.1.1 Definición de preguntas de investigación

Se definen las preguntas de investigación acorde con el tema propuesto y lo que se requiere conocer acerca de la detección de cáncer de pulmón mediante aprendizaje automático.

¿Cuáles son los conjuntos de datos que se emplean en trabajos de investigación relacionados a la clasificación de cáncer de pulmón

utilizando imágenes tomográficas?

La primera cuestión permitirá identificar los diversos conjuntos de datos empleados encada artículo de investigación, indicando si trabajan con datos, imágenes, número de pacientes, tamaño del conjunto de datos, ordenado por tipos de cáncer o clasificado como maligno o benigno.

¿Cuáles son los atributos de los data set de imágenes tomográficas que se toman en consideración para la clasificación del cáncer de pulmón?

La segunda pregunta de investigación busca identificar los atributos utilizados generalmente para este tipo de cáncer, de igual manera esto dependerá del conjunto de datos que seleccionen los investigadores.

¿Cuáles son métodos que se utilizan con mayor frecuencia para la clasificación del cáncer de pulmón utilizando imágenes tomográficas?

Esta pregunta permite a identificar el detalle de la metodología propuesta mediante un gráfico, solo unos poco detallan su diagrama con las etapas y sub etapas del proceso de desarrollo de la investigación.

¿Cuáles son las métricas que se utilizan para medir los modelos?

Generalmente el detalle de cada uno no es mencionado solo de los más usuales como sensibilidad, F1, AUC entre otros existen otras meticas aparte de ellas, algunas investigaciones consideran la descripción de cada indicador como su expresión, significado y detalle de formula en la etapa de evaluación.

¿Cuáles son los trabajos de investigación que han obtenido mejores resultados en la clasificación de cáncer de pulmón utilizando imágenes tomográficas?

La mayoría de las métricas como la precisión, exactitud, sensibilidad, y otros si logran pasar la balla del 90% pero cuando la investigación considera más indicadores oscilan entre 80%.

3.3.1.2 Palabras clave y sinónimos

En esta sección definimos acorde con nuestras preguntas de investigación y el titulo las palabras claves con sus sinónimos que posteriormente nos apoyara en la

generación de la cadena de búsqueda.

Tabla 7. Palabras clave y sinónimos.

Palabra clave	Sinónimo	Relación
Detección del sistema	Aplicación de escritorio, aplicaciones web	Salida
Cáncer	Adenocarcinoma de pulmón, cáncer de pulmón	Intervención
Comparación	Métricas de evaluación	Comparación
Detección de cáncer de pulmón	Diagnóstico de nódulos pulmonares, reconocimiento de nódulos	Intervención
Aprendizaje automático	Algoritmos, inteligencia artificial, clasificación, técnicas	Población
Metodología	Nuevas metodologías, metodologías habituales	Comparación

3.3.1.3 Definición de fuentes

Se selecciona las bases de datos en donde se ingresa la cadena de búsqueda personalizada para cada una puesto que no todas poseen los mismos parámetros en búsqueda avanzada.

Tabla 8. Fuente y cadena de búsqueda.

Fuente	Cadena
ACM	[[All: "machine learnig"] OR [All: "algorithms"] OR [All: "artificial intelligence"] OR [All: "classification"] OR [All: "techniques"]] AND [[All: "cancer"] OR [All: "lung adenocarcinoma"] OR [All: "cancer lung"] OR [All: "lung cancer detection"] OR [All: "diagnosis of pulmonary nodules"] OR [All: "nodule recognition"]] AND [[All: "methodology"] OR [All: "new methodologies"] OR [All: "usual methodologies"]] AND [[All: "system detection"] OR [All: "desktop application"] OR [All: "web applications"] OR [All: "comparison"] OR [All: "evaluation metrics"]]
IEEE	("machine learnig" OR "algorithms" OR "artificial intelligence" OR "classification" OR "techniques") AND ("cancer" OR "Lung adenocarcinoma" OR "cancer lung" OR "lung cancer

	detection" OR "Diagnosis of pulmonary nodules" OR "nodule recognition") AND ("methodology" OR "new methodologies" OR "usual methodologies")
Science Direct	("machine learnig" OR "algorithms" OR "classification" OR "detection" OR "techniques") AND ("Lung adenocarcinoma" OR “lung cancer detection" OR "Diagnosis of pulmonary nodules" OR "nodule recognition")

3.3.1.4 Criterios de inclusión y exclusión

Posteriormente los criterios definidos serán relevantes en la calificación de las investigaciones.

Tabla 9. Criterios de inclusión y exclusión.

Inclusión	Exclusión
Artículos de investigación	Artículos no relacionados al aprendizaje automático
Artículos relacionados a aprendizaje automático	Artículos de revisión
Artículos que poseen introducción, problema, método, resultado, conclusiones y trabajos futuros	Artículos no relacionados al cáncer de pulmón
Artículos de hace 6 años	Artículos de hace 7 años
Publicaciones en ingles	Publicaciones en español

3.3.1.5 Extracción de información

En esta sección se colocará información relevante respecto a cada investigación como autores, fuente, año, resumen, procedencia, país entre otros con la finalidad de identificar claramente los datos generales con los cuales se realiza la investigación.

Tabla 10. Extracción de datos.

Elemento	Descripción
Autores	Autores del articulo

Fuente	Tipo de publicación
Titulo	Título del articulo
Año	Año de publicación del articulo
País	País de afiliación de los autores
Procedencia	Fuente de publicación
Tipo de articulo	Basado en la metodología propuesta del articulo
Ámbito	Ámbito de aplicación del artículo

3.3.1.6 Realización de la investigación

Después de definir una cadena de búsqueda acorde con cada fuente de información se seleccionarán y se guardaran en archivo bibtext que se emplearan en la siguiente sección.

3.3.1.7 Importar estudios

Los archivos guardados en formato bibtext en el paso anterior se agregaron en su respectiva base de datos, luego aplicaremos el primero filtro rechazando, eliminando, o identificando artículos duplicados, solo aquellos que su estado este en aceptado pasaran a la etapa final de calificación de artículos.

3.3.1.8 Calificar estudios

Finalmente se comienza a evaluar cada artículo de manera individual pasando por cada pregunta y calificaron de manera objetiva los 115 trabajos de investigación, solo se aceptaron 66.

3.3.1.9 Resultados de investigación previa

En este apartado se presentan los artículos que pasaron los filtros de calidad, después del protocolo mencionado precedentemente se obtuvo un total de 357 artículos de las bases de datos Science Direct, IEEE y ACM mostradas en la figura 29. Luego de la primera fase se seleccionaron artículos por título quedando 48, 48 y 19 respectivamente por cada conjunto de datos, también se aprecia un total de 238 artículos rechazados y 4 artículos duplicados, aquellos excluidos fueron los que excedieron el tiempo límite de más de cinco años, artículos que no sean de investigación, sin relación al cáncer de pulmón y aprendizaje automático. Del porcentaje aceptado en su resumen la mayoría contemplaba una metodología en

relación al contexto, pero en la evaluación de calidad a través de preguntas no superaron el alcance establecido de 3.0 y un límite de 5.0 con un valor de 1.0 para cada incógnita esto resulto en la selección de 66 artículos para la inclusión de este estudio, se revisaron en su totalidad durante la última etapa, en adelante se muestra la lista de los elegidos y los elementos extraídos.

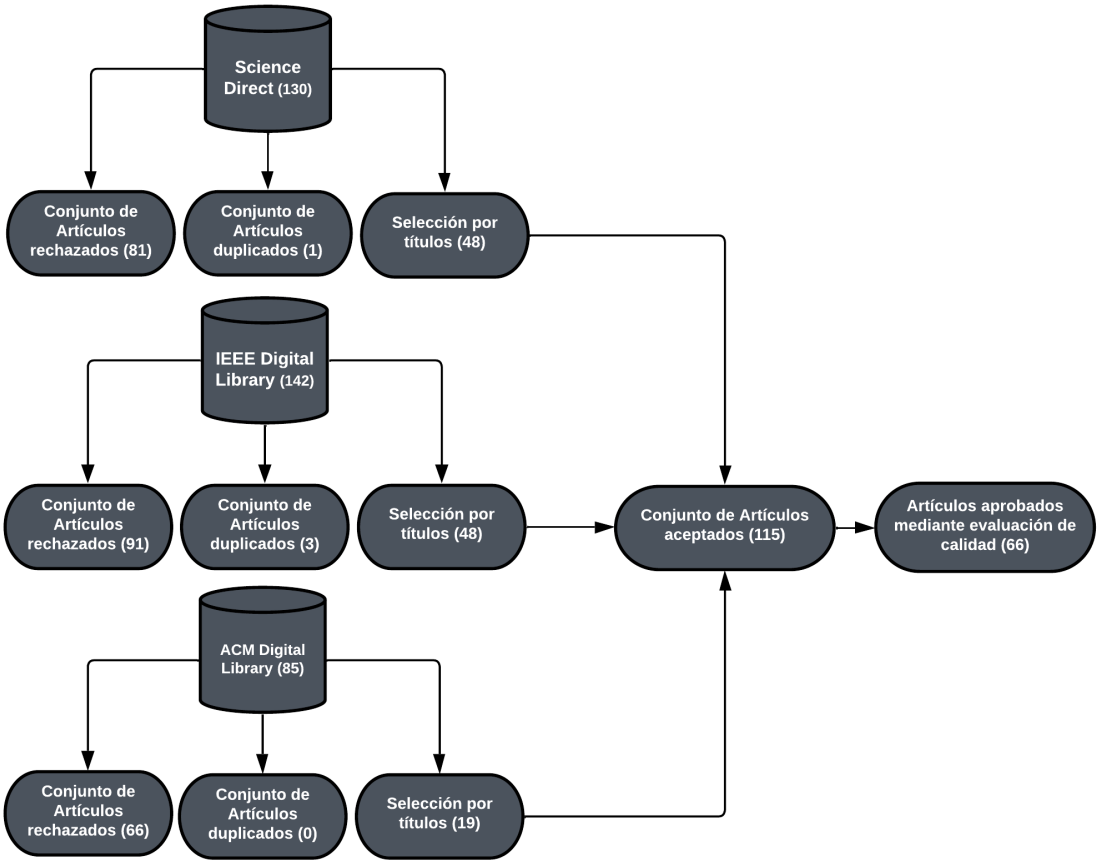


Figura 29. Publicaciones por base de datos

Tabla 11. Lista de artículos seleccionados.

#	Referencia de Art.	Año	Fuente	Ámbito de la aplicación
1	[32]	2019	Conferencia	Medicina Personalizada
2	[33]	2020	Conferencia	Clasificaron/detección mediante imágenes
3	[34]	2020	Conferencia	Clasificaron/detección mediante imágenes
4	[35]	2019	Conferencia	Clasificaron/detección mediante imágenes
5	[36]	2020	Conferencia	Clasificaron/detección mediante imágenes
6	[37]	2020	Conferencia	Clasificación/detección mediante datos
7	[38]	2020	Conferencia	Clasificación/detección mediante datos
8	[39]	2020	Conferencia	Clasificaron/detección mediante imágenes
9	[40]	2021	Conferencia	Clasificación/detección mediante datos
10	[41]	2021	Conferencia	Clasificaron/detección mediante imágenes
11	[42]	2018	Conferencia	Clasificaron/detección mediante imágenes
12	[43]	2018	Conferencia	Clasificaron/detección mediante imágenes

13	[44]	2018	Simposio	Clasificación y detección mediante datos
14	[45]	2018	Conferencia	Clasificaron/detección mediante imágenes
15	[46]	2018	Conferencia	Clasificaron/detección mediante imágenes
16	[47]	2020	Revista	Clasificaron/detección mediante imágenes
17	[48]	2019	Revista	Clasificación/detección mediante datos
18	[49]	2020	Revista	Clasificaron/detección mediante imágenes
19	[11]	2021	Revista	Clasificaron/detección mediante imágenes
20	[4]	2018	Revista	Clasificaron/detección mediante imágenes
21	[50]	2019	Revista	Clasificaron/detección mediante imágenes
22	[51]	2020	Revista	Clasificaron/detección mediante imágenes
23	[52]	2020	Revista	Clasificaron/detección mediante imágenes
24	[14]	2021	Conferencia	Clasificaron/detección mediante imágenes
25	[53]	2020	Revista	Clasificaron/detección mediante imágenes
26	[54]	2018	Conferencia	Clasificaron/detección mediante imágenes
27	[7]	2020	Conferencia	Clasificaron/detección mediante imágenes
28	[55]	2018	Conferencia	Clasificaron/detección mediante imágenes
29	[6]	2020	Conferencia	Clasificaron/detección mediante imágenes
30	[56]	2020	Revista	Clasificaron/detección mediante imágenes
31	[57]	2019	Conferencia	Clasificaron/detección mediante imágenes
32	[58]	2020	Revista	Clasificaron/detección mediante imágenes
33	[30]	2020	Revista	Clasificaron/detección mediante imágenes
34	[59]	2019	Conferencia	Clasificaron/detección mediante imágenes
35	[60]	2017	Simposio	Clasificaron/detección mediante imágenes
36	[61]	2019	Conferencia	Clasificaron/detección mediante imágenes
37	[62]	2019	Conferencia	Clasificaron/detección mediante imágenes
38	[63]	2020	Revista	Clasificaron/detección mediante imágenes
39	[64]	2020	Revista	Clasificaron/detección mediante imágenes
40	[8]	2020	Conferencia	Clasificaron/detección mediante imágenes
41	[65]	2020	Conferencia	Clasificaron/detección mediante imágenes
42	[66]	2020	Conferencia	Clasificaron/detección mediante imágenes
43	[31]	2019	Conferencia	Clasificaron/detección mediante imágenes
44	[5]	2020	Revista	Clasificaron/detección mediante imágenes
45	[67]	2019	Revista	Clasificaron/detección mediante imágenes
46	[18]	2019	Conferencia	Clasificaron/detección mediante imágenes
47	[16]	2017	Conferencia	Clasificaron/detección mediante imágenes
48	[12]	2020	Conferencia	Clasificaron/detección mediante imágenes
49	[13]	2020	Conferencia	Clasificaron/detección mediante imágenes
50	[9]	2020	Conferencia	Clasificaron/detección mediante imágenes
51	[10]	2020	Conferencia	Clasificaron/detección mediante imágenes
52	[68]	2021	Conferencia	Clasificaron/detección mediante imágenes
53	[69]	2022	Revista	Clasificaron/detección mediante imágenes
54	[70]	2020	Revista	Clasificaron/detección mediante imágenes
55	[71]	2021	Revista	Clasificaron/detección mediante imágenes
56	[72]	2021	Revista	Clasificaron/detección mediante imágenes
57	[73]	2021	Revista	Clasificaron/detección mediante imágenes
58	[74]	2021	Revista	Clasificaron/detección mediante imágenes
59	[75]	2022	Revista	Clasificaron/detección mediante imágenes
60	[76]	2021	Revista	Clasificaron/detección mediante imágenes
61	[77]	2021	Revista	Clasificaron/detección mediante imágenes
62	[78]	2021	Revista	Clasificaron/detección mediante imágenes
63	[79]	2021	Revista	Clasificaron/detección mediante imágenes
64	[80]	2021	Revista	Clasificaron/detección mediante imágenes
65	[81]	2020	Revista	Clasificaron/detección mediante imágenes

66	[82]	2021	Revista	Clasificaron/detección mediante imágenes

Información básica sobre los artículos

En este apartado se realiza un análisis sobre los resultados obtenidos referente a los elementos mencionados en la lista de extracción de información.

Año de publicación, procedencia y distribución geográfica

El límite de tiempo de nuestro protocolo de investigación comienza en 2017 hasta el 2022. La figura 30 se muestra la cantidad de publicaciones por año de los artículos seleccionados, la mayor parte de la selección se publicó en el 2020, después 15 seleccionados en el 2019 mientras que en el 2021 y 2022 representa solo el 27% de la elección con unos 18 artículos en total, por último, solo se seleccionaron 5 investigación del 2018 y 3 artículo del 2017. El aumento en el número de investigaciones representa el interés del área de estudio asimismo se puede distinguir una caída en el número de publicaciones del presente año debido a la no disponibilidad de los mismos.

En la figura 31 es una representación porcentual de la procedencia de cada artículo, se idéntica que gran parte de las investigaciones son producto de diversos centros de académicos con un 57.58%, el porcentaje de industria solo simboliza 7.58% y 34.85% fueron escritos en ambos sectores.

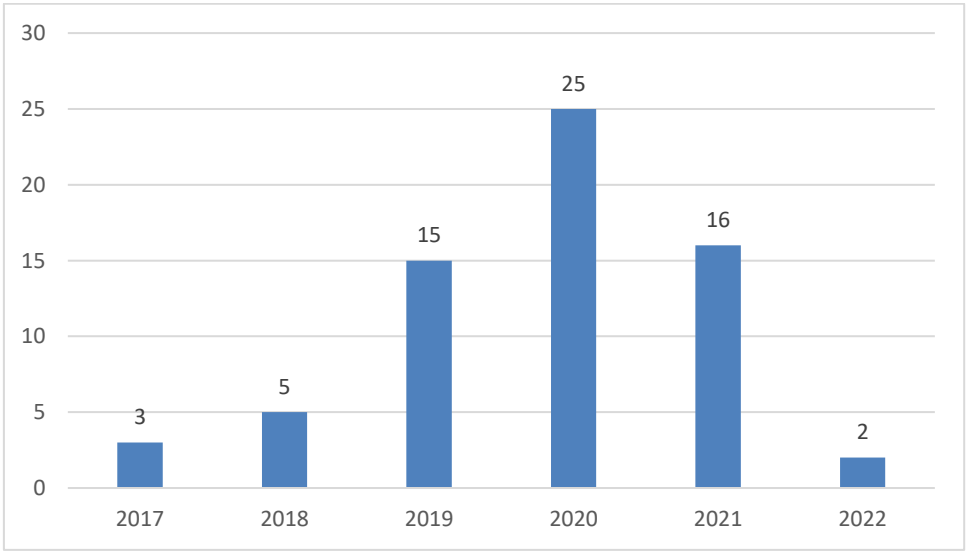


Figura 30. Publicaciones por año de los artículos.

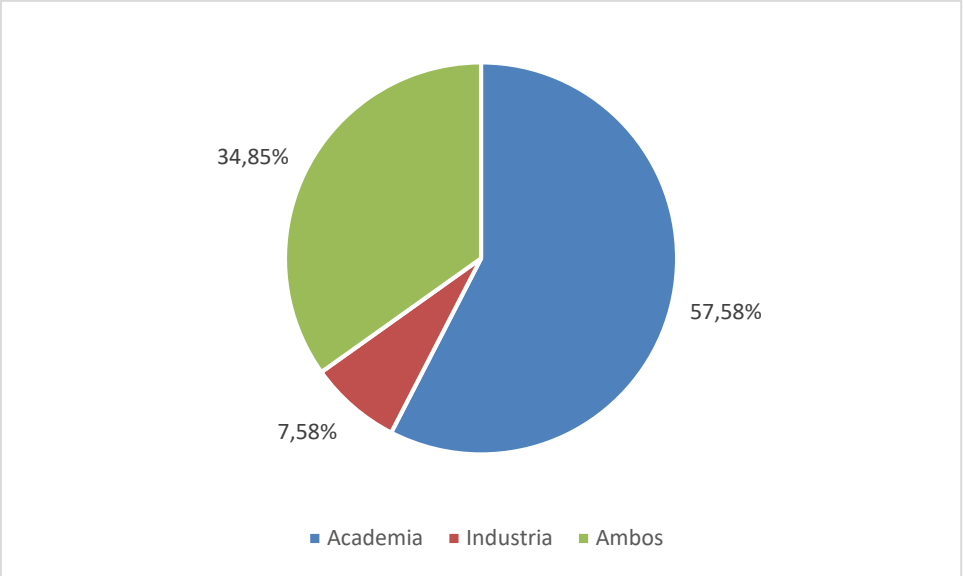


Figura 31. Procedencia de los artículos.

Por ultimo para aclarar la distribución geográfica de los involucrados sobre la aplicación del aprendizaje automático en la detención del cáncer de pulmón empleamos el dato de extracción “país” acorde con la afiliación de los autores de cada investigación seleccionada se muestran el número de artículos que realizaron publicaciones referente al tema , en la figura 32 se aprecia que los primeros seis países aportaron una investigación cada uno luego Grecia, Malasia y Arabia saudita 6 en total, Bangladesh y Estados Unidos representando un 12.12% con 8 artículos seleccionados cada uno, Finalmente los dos últimos países en el grafico representan el mayor porcentaje de contribución en la investigación, China (diecinueve artículos) e India (veinte cuatro artículos) con 28.79% y 36.36% respectivamente

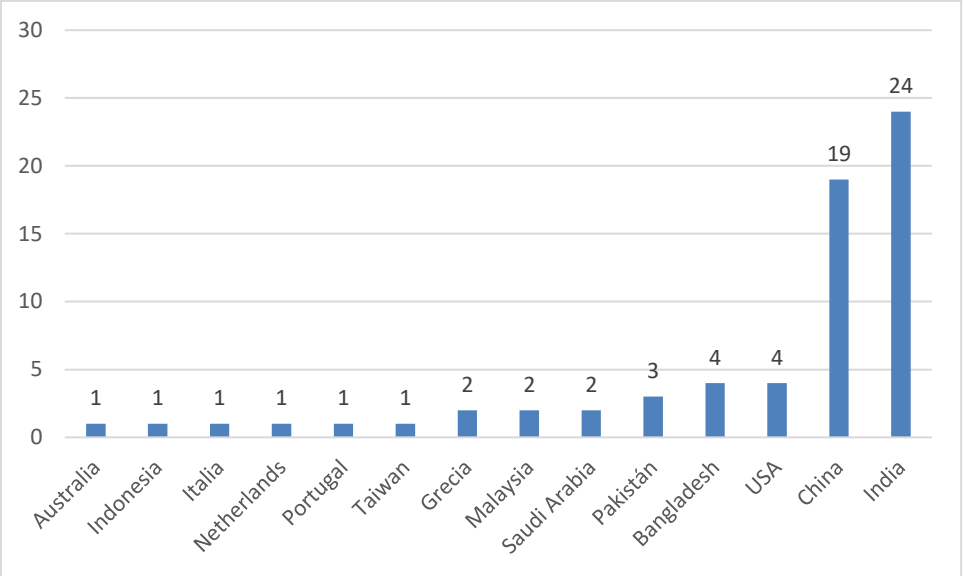


Figura 32. Distribución de artículos por país de autores

Tipo de fuente

Esta sección muestra el número de artículos elegidos por el estudio dependiendo del lugar de publicación si es en una conferencia, revista y simposio en el caso de los seleccionados, también se pueden encontrar en capítulos de libros o talleres. En la figura 33 se muestra el número de investigaciones seleccionadas, de los 66 artículos elegidos el 3.03% fueron publicados en Simposios (2 artículos), 45.45% de revistas (30 artículos) y 51.52% de conferencias (34 artículos).

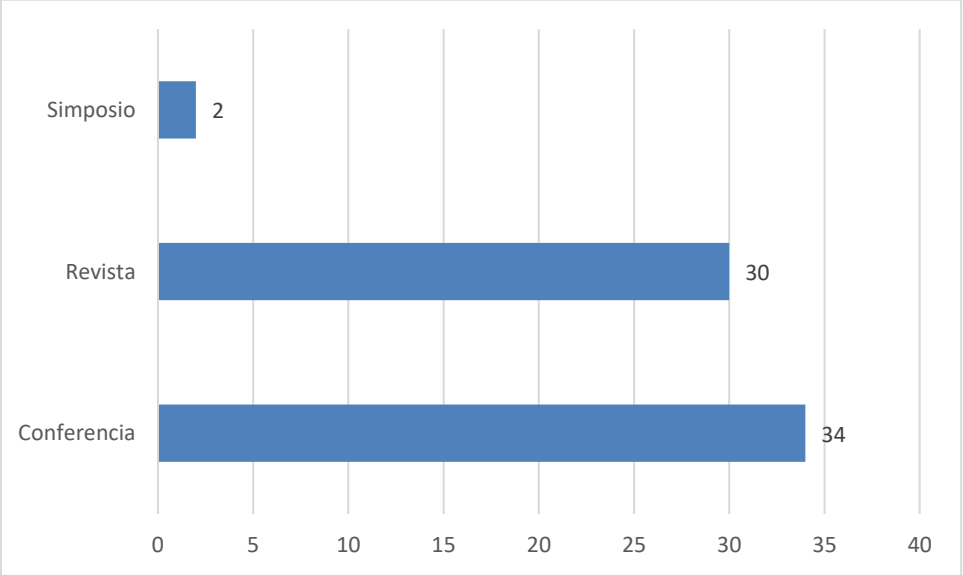


Figura 33. Distribución de artículos por tipo de fuente

Clasificación de los trabajos seleccionados

En la figura 34 mostramos la porción de investigaciones elegidas que son artículos técnicos y documentos aplicados al aprendizaje automático, se puede observar que solo el 3.03% son artículos de aplicación y el 96.97% son técnicos.

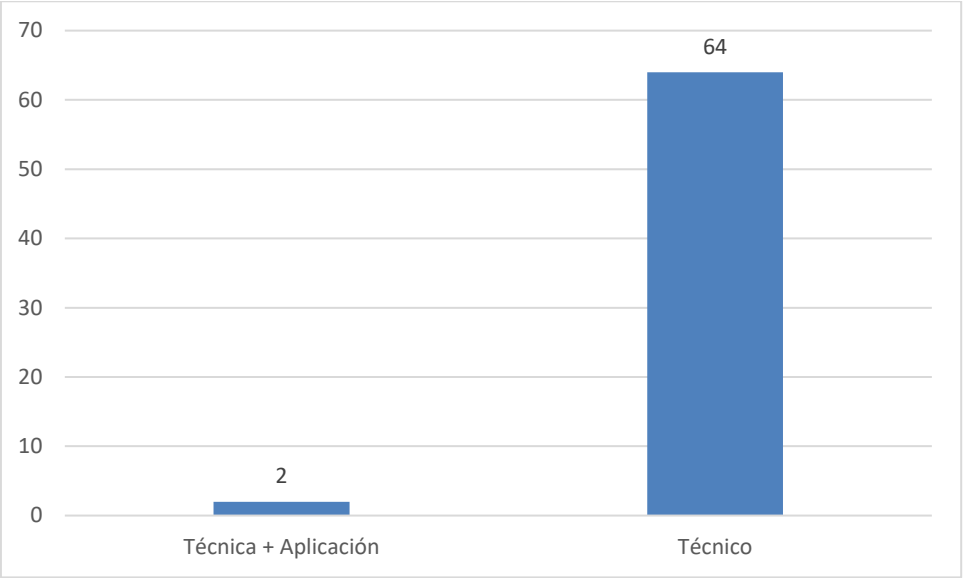


Figura 34. Tipo de artículos

En la siguiente figura 35 se evidencia a detalle la cantidad de investigaciones por año según su tipo a partir del 2017 hasta el 2019 solo se publicaron documentos técnicos, en adelante se comienzan a publicar artículos aplicados aun que son muy escasos y algunos no cumplieron con los criterios de calidad se excluyeron del estudio, se puede identificar el creciente crecimiento de investigaciones relacionadas al tema.

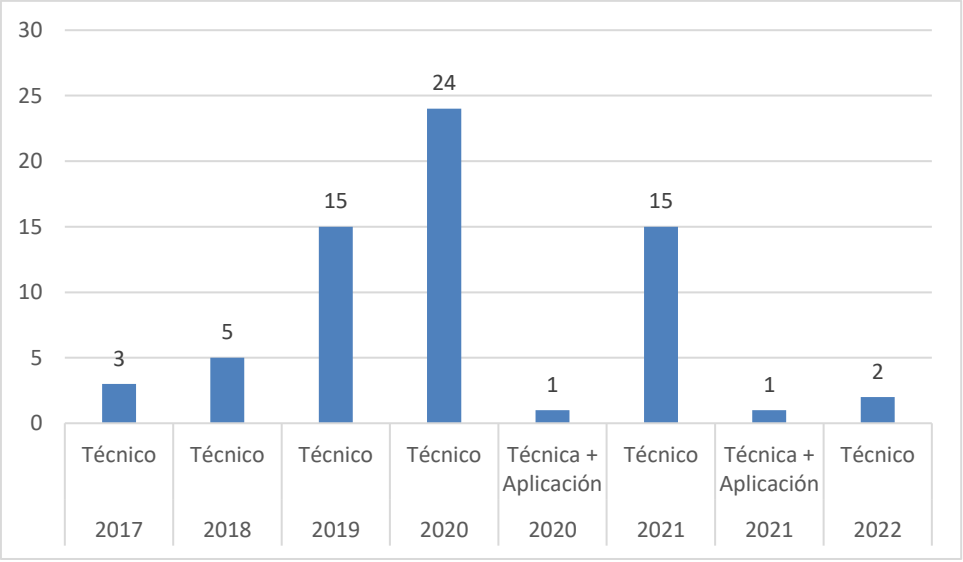


Figura 35. Cantidad de publicaciones por año

Tabla 12. Ámbito de la aplicación.

Ámbito de la aplicación	Artículos
Medicina Personalizada	[32]
Clasificación/detección mediante datos	[37][38][40][48]
Clasificación/detección mediante imágenes	[32-38,40-46,48-81]

Fuente: Elaboración propia

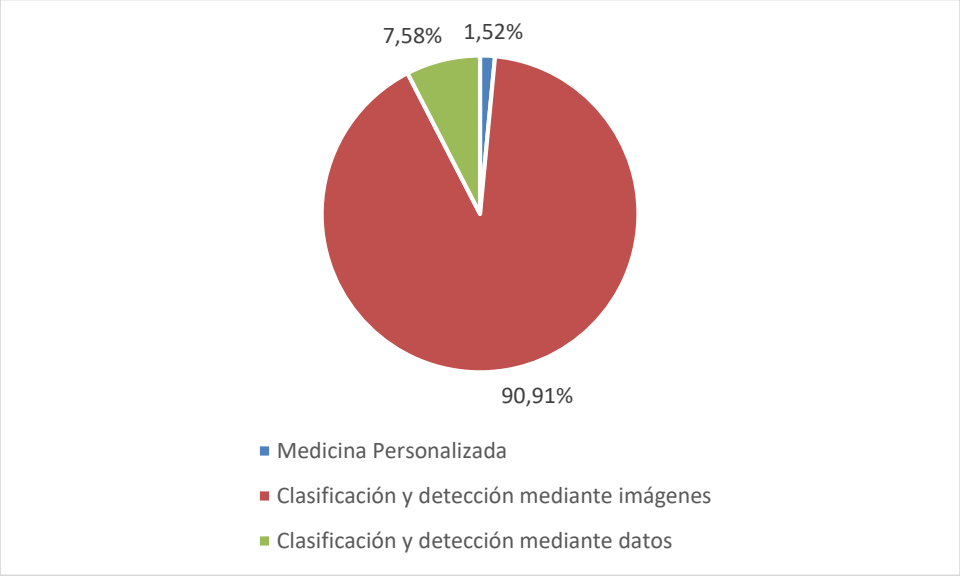


Figura 36. Porcentaje de distribución por ámbito

Medicina Personalizada

De las investigaciones calificadas que superaron el umbral propuesto solo una de ellas hace referencia a el primer ámbito, esta pronostica el cáncer y la medicina personalizada. En [32] para tratar la enfermedad se han tratado terapias, quimioterapias, inmunoterapias entre otros no obstante los pacientes responden diferente debido a su genética, estado de salud, etnia y tratamientos anteriores.

Clasificación/detección mediante datos

Este ámbito consiste simplemente en aquellos estudios que consideran en la etapa de selección de datos aquellos conjuntos que contiene características como en [37][38][40][48]

Clasificación/detección mediante imágenes

De manera contraria al mencionado anteriormente este ámbito consiste en el resto de estudios que representa la mayor parte de los seleccionados en donde se seleccionad bases de datos que contiene imágenes del cáncer de pulmón y que posteriormente se extraen características de las mismas para posteriormente usarlas en los algoritmos sea un clasificador o arquitectura de red neuronal.

3.3.1.10 Discusión de investigación previa

En siguiente apartado consiste en contra examinar las interrogantes planteadas por esta investigación a través de los resultados del estudio.

¿Cuáles son los conjuntos de datos que se emplean en trabajos de investigación relacionados a la clasificación de cáncer de pulmón utilizando imágenes tomográficas?

Los conjuntos de datos empleados son aquellos que se mencionan en la tabla 13. Detallando el nombre de la base de datos, el tipo de imagen que contiene, el número de participantes, la referencia a los artículos, numero de imágenes, tamaño de descarga y el contador del total de uso de cada uno, en resumen, se puede apreciar que el conjunto de datos más empleado es LIDC-IDRI, en segundo lugar, LUNA16, tercero Data Science Bowl 2017, cuarto SPIE-AAPM Lung CT Challenge, quinto LC25000 con un contador total de numero de uso de respectivamente 17, 10, 3, 2 y 2.

Tabla 13. Conjuntos de datos usados en los artículos de investigación.

Conjunto de datos	Tipo y link de acceso	Tipo de img	Participantes	Artículos	Numero de img	Tamaño de img GB	Total
the Cancer Genome Atlas Database (TCGA)	Publico https://portal.gdc.cancer.gov/		12000	[32]			1
LC25000	Publico https://arxiv.org/abs/1912.12142v1			[33][69]	15000 pulmón y 1000 de colon	1.89	2
LUNA16	Publico https://luna16.grand-challenge.org/Data/	Tomografía computarizada CT		[34][49][55][56][61][35][33][72][75][76][78]	888		10
DataSet	Privado	resonancia magnética RM	89	[35]			1
LIDC-IDRI	Publico https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI	Tomografía computarizada CT, radiografía computarizada CR y Rayos x digitales DX	1010	[36][4], [46][4][53][55][6][57][60][64][66][5][18][70][71][72][80][34][37]	244527	124	17

GSE4115, GSE33356, GSE3141, GSE8894, and GSE40419	Publico https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi		187,120,111, 138,164	[37]		1
DataSet	Privado			[38]	100	1
CSIRO	Privado	Radiografía de tórax		[39]	153	1
BRCA, LUAD, LUSC, KICH, KIRC, KIRP, BLCA, COAD, ESCA, HNSC, LIHC, PRAD, STAD, THCA y UCEC	Publico https://xenabrowser.net/datapages/			[40]		1
DataSet	Privado	Tomografía computarizada CT		[41]		1
DataSet	Privado	Tomografía computarizada CT y tomografía por emisión de positrones (PET)	20	[42]	1000	1

DataSet	Privado	Tomografía computarizada de dosis baja LDCT		[43]		1
Lung Cancer Disease dataset	Publico https://archive.ics.uci.edu/ml/datasets/lung+cancer			[44]		1
DataSet of Dr. Vithalrao Vikhe Patil Medical College and Hospital	Privado			[45]	322	1
ELCAP y LIDC- IDRI	Publico http://www.via.cornell.edu/lungdb.html	Tomografía computarizada CT	130	[47][73]		1
DataSet	Privado			[48]		1
DataSet	Privado	Tomografía computarizada CT	1222	[11]		1

TCGA-LUAD	Publico tps://wiki.cancerimaging chive.net/display/Public/ CGA-LUAD	Tomografía computarizada CT, Terapia física PT y Medicina Nuclear NM	69	[50]	48931	18.3	1
DataSet	Privado	Tomografía computarizada CT	501	[51]			1
DataSet	Privado	Tomografía computarizada CT		[52]			1
DataSet Chest CT Scan images	Publico https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images	Tomografía computarizada CT		[14]	1000	124.96 MB	1
Clinical dataset from Shanghai Sixth People's Hospital (SPH)	Privado	Tomografía computarizada CT		[53]			1
Lung CT scan dataset from an oncology center in Mangalore, Karnataka	Privado	Tomografía computarizada CT		[54]			1

Instituto Médico y Dental Avanzado (AMDI)	Privado	Tomografía computarizada CT		[7]	100		1
DataSet de Shandong Provincial Hospital	Privado	Tomografía computarizada CT		[58]	1222		1
Imágenes médicas avanzadas del Instituto Nacional del Cáncer (NCI) y la base de datos de imágenes de la Universidad de Washington y base de datos de Kaggle	Privado los 2 primeros conjuntos y Público el tercer dataset https://www.kaggle.com/datasets/kmader/lungnodemalignancy	Tomografía computarizada CT		[30]			1
Conjunto de base de datos de the Cancer Genome Atlas Database (TCGA)	Publico https://portal.gdc.cancer.gov/			[59]	1000		1
Data Science Bowl 2017	Publico https://www.kaggle.com/competitions/data-science-bowl-2017/data	Tomografía computarizada CT		[55][62][16]			3
SPIE-AAPM Lung CT Challenge	Publico https://wiki.cancerimagi	Tomografía computarizada CT	70	[63][8]	22489	12.1	2

	ngarchive.net/display/Public/SPIE-AAPM+Lung+CT+Challenge					
DataSet Departamento de Oncología, Hospitales Manipal	Privado	Tomografía computarizada de hueso, cerebro, pulmón, riñón y cuello		[65]	500	1
DataSet	Publico https://www.cancerimagingarchive.net/	Tomografía computarizada CT	l	[31]	4682	1
FAH-GMU	Privado	Tomografía computarizada CT	l5	[5]		1
DataSet collected from Cancer imaging Archive (CIA) dataset	Público https://www.cancerimagingarchive.net/	Tomografía computarizada CT		[67]	5043	1
DataSet	Privado	Tomografía computarizada CT		[12]		1

DataSet	Privado	Tomografía computarizada CT		[13]	78	1
DataSet collected from the Rajiv Gandhi Cancer Institute and Research Centre	Privado	Tomografía computarizada CT		[9]	300	1
DataSet	Privado	Tomografía computarizada CT		[10]		1
DataSet Origi Med Inc	Privado		55	[68]	155	1
DataSet	Privado	Tomografía computarizada CT		[74]		1
LUNA16, LNDb, ILCID	Público y el resto privados https://luna16.grand-challenge.org/Data/	Tomografía computarizada CT	2361	[77]		1
DataSet	Privado	Tomografía computarizada CT	920	[81]	920	1
DataSet	Privado	Tomografía computarizada CT	247	[82]	247	1

¿Cuáles son los atributos de los data set de imágenes tomográficas que se toman en consideración para la clasificación del cáncer de pulmón?

Las redes neuronales en los últimos años han emergido demasiado rápido tanto así que es difícil encontrar artículos de investigaciones que hagan uso de datos en lugar de imágenes para detectar enfermedades cancerígenas, algunos de los atributos que se emplearon en [52] son Edad, sexo (M/F), fuma (S/N), Estadio (IA1, IA2, IA3, IB), Invasión pleural, Grado Histológico (A/M/B), EGFR que representan características patológicas relacionas a esta enfermedad. Los atributos que se consideran en [12] en la sección de extracción de características de imágenes son área de cáncer, contraste, energía, entropía y homogeneidad. Actualmente se hace uno de conjunto de datos libres o exclusivos como Kaggle Data Science Bowl 2017, LIDC-IDRI los mencionados en [55], LUNA16 en [49] Delhi en [9], imágenes microscópicas del pulmón de la facultad de medicina “Dr. Vithalrao Vikhe Patil” en [45] entre otros.

¿Cuáles son métodos que se utilizan con mayor frecuencia para la clasificación del cáncer de pulmón utilizando imágenes tomográficas?

El método es importante debido a que detalla cómo está estructurado el trabajo generalmente optan por trabajar con un conjunto de datos sea público o privado que contenga imágenes y otros pocos con características, dependiendo de ello se trabajara primero mediante técnicas de procesamiento de imágenes después se aplicara los algoritmos de aprendizaje automático: redes neuronales, máquina de vectores de soporte, K vecinos más próximos entre otros.

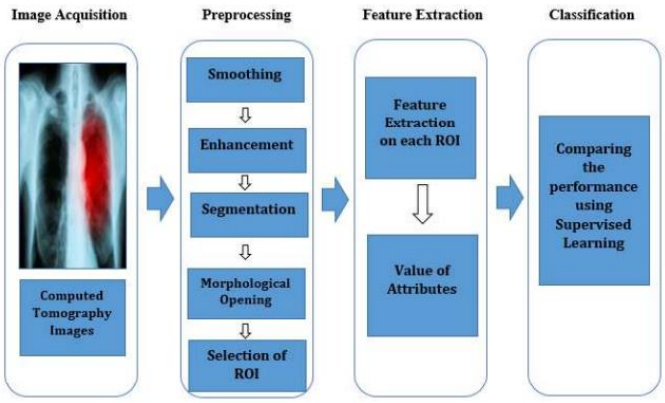


Figura 37. Diagrama de trabajo de [31]

Comúnmente no se suele detallar a profundidad el proceso de la investigación en el flujo de trabajo como en [31] especifica de manera clara su propuesta mediante el diagrama propuesto constando de 4 etapas en la primera se

adquieren las imágenes médicas luego pasaron al preprocesamiento donde se aplica suavizado, mejora, segmentación, selección de región de interés (ROI) luego se extraen características en cada ROI y en la etapa final de clasificación se compara el rendimiento de algoritmos de aprendizaje supervisado SVM, KNN, entre otros.

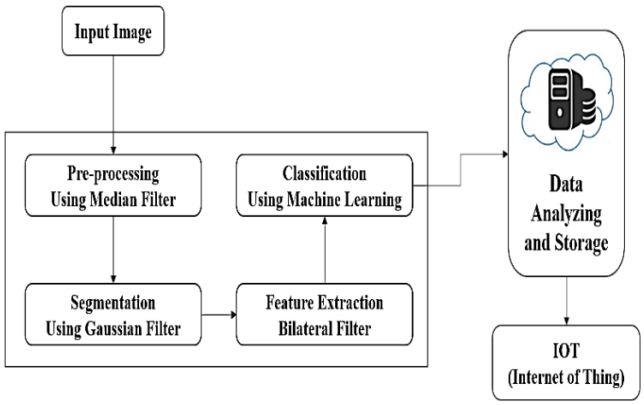


Figura 38. Diagrama de trabajo de [41]

De igual forma en [41] se propone una metodología similar, pero emplean diferentes filtros en sus etapas ya sea para identificar la forma, tamaño y análisis del resultado a ello se suma el desarrollo de un software en el que se ingresa la imagen, se analiza y muestra el resultado si se posee la enfermedad o no.

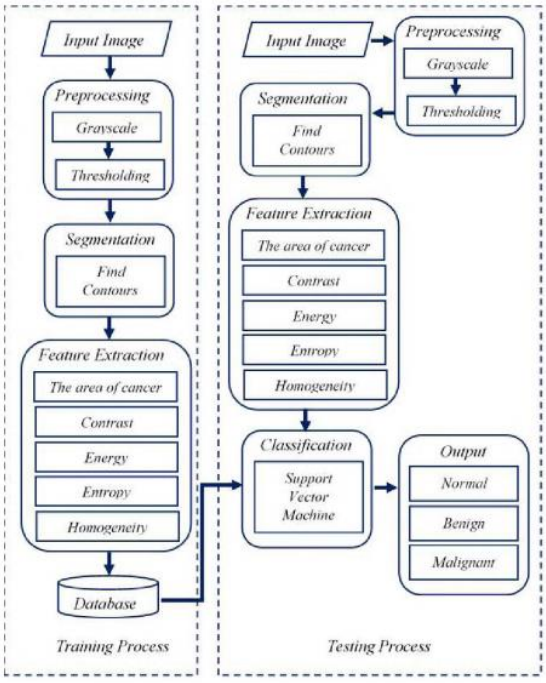


Figura 39. Diagrama de trabajo de [12]

Por otro lado, en [12] la metodología propuesta es bastante detallada este divide

sus procesos en dos partes entrenamiento y prueba, en la primera entran la imágenes, luego en el procesamiento se realiza una escala de grises y umbral, en la segmentación se separa el objeto del resto, después en la extracción de características se considera el área del cáncer, contraste, entropía, energía y homogeneidad, finalmente en la última etapa de clasificación se detalla dicho proceso y las salidas del cáncer: normal, benigna y maligna a pesar de ello el diagnóstico del modelo es de 83%.

Las etapas generalmente consideradas son primero adquirir o construir de un conjunto de datos, posteriormente podemos preprocesar los datos o imágenes, extracción de región de interés o características y finalmente a la etapa de clasificación podemos desarrollar un sistema de predicción en base a los resultados obtenidos esto depende del enfoque que tenga la investigación.

El estudio no considero artículo de revisión sistemática de la literatura en relación al cáncer de pulmón aplicando aprendizaje automático al contrario solo aceptamos de investigación por lo que se ejecuta la propuesta de cada estudio, se precisa la metodología a través de una gráfica y se detalla en la sección de metodología, de los seleccionados solo dos desarrollaron software ambos estudios [41] [9] de equivalente procedimiento primero ingresamos la imagen, ejecutamos y muestra la categorización positiva o negativa del cáncer.

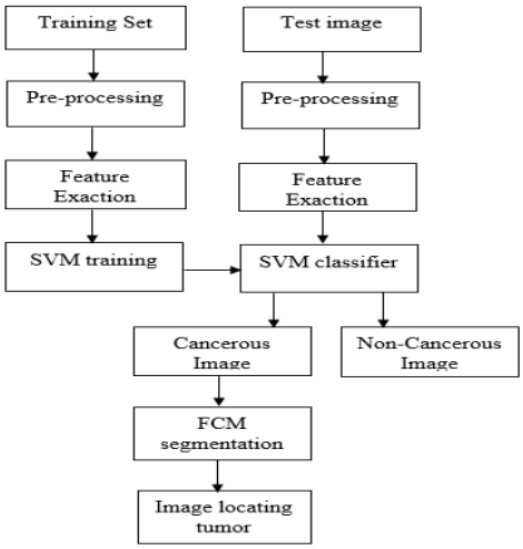


Figura 40. Diagrama de trabajo de [54]

Los modelos propuesto técnicos y aplicados pueden dividirse de manera

genérica en cuatro etapas: preprocesamiento de imágenes, segmentación, extraer características y clasificación que son las más usuales en la detención del cáncer de pulmón en el caso de [54] fragmenta el conjunto de datos en entrenamiento y prueba, hace uso de las etapas mencionadas anteriormente y emplea SVM como clasificador dando como resultado una respuesta negativa o afirmativa, si es positiva localiza en la imagen el tumor y se muestra. Primeramente se adquiere un conjunto de datos de un centro oncológico con imágenes en extensión dicom y un grosor de 2.5 mm en la fase de preprocesamiento aquellas imágenes que contienen ruido u otras fallas se aplica ecualización de histograma para optimizar el contraste de las mismas, y para la calidad y restricción de amplificación del ruido se empleó CLACHE, después en la extracción de características se utiliza la matriz de coocurrencia de nivel de gris para las características, los atributos generados de las imágenes fueron energía, correlación, contraste, disimilitud, varianza y probabilidad máxima. En la clasificación seleccionaron SVM inicialmente se preparó con el conjunto de datos de entrenamiento y se evaluó con los datos de prueba, después muestra la salida cancerosa o no, posteriormente en la segmentación aquellas diagnosticadas como positivas para el cáncer se aplica la agrupación media difusa C para mostrar en área cancerosa en la imagen.

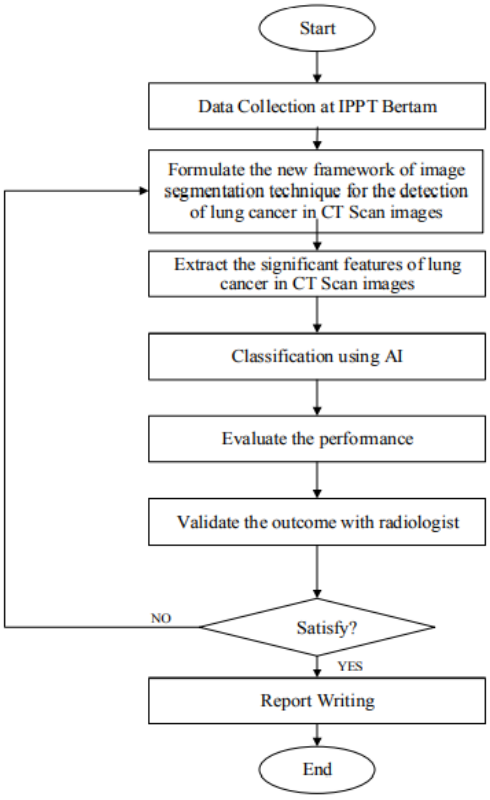


Figura 41. Diagrama de trabajo de [7]

El articulo [7] plantea un proceso similar primero recolecta datos obtenidos de la base de datos de AMDI con un total de 100 instancias en formado dicom y una dimensión de 512x512 pixeles, segundo en el procesamiento de los mismos se aplica el filtro mediano para eliminar el ruido, tercero se segmenta manualmente haciendo uso de ImageJ invirtiendo los colores luego se transforma a binario y en Photoshop se cargan para quitar el fondo de las imágenes, cuarto se extrae características geométricas como centroide, perímetro y área para medir la dimensionalidad física y finaliza con el clasificador KNN escogido para identificar el estadio del cáncer y en base a ello los médicos expertos seleccionar un tratamiento adecuado.

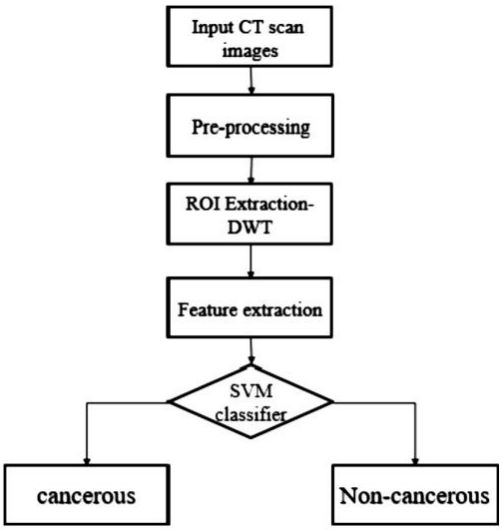


Figura 42. Diagrama de trabajo de [47]

La investigación [47] considera como paso inicial la entrada de las tomografías, pasando al preprocesamiento, región de interés, extracción de características y usar el clasificador SVM. El conjunto de datos este compuesto de 130 ejemplares 84 de LIDC y 46 de ELCAP que son otras bases de datos, debido a que las tomografías computarizadas poseen ruido se aplica el filtro de “sal y pimienta” para eliminarlo, también se sumó a ello el filtro Weiner para reducir el nivel de ruido y excluir el ruido gaussiano. El ROI extrae pixeles que cambian al dominio transformado, se seleccionaron las transformadas de ondículas (db1, db2 y db4), al aplicar una transformada en sub-bandas de frecuencia (LL1, LH1, HL1 y HH1) se logra el análisis multirresolución. En la siguiente etapa se extraen un conjunto de 264 características a partide de la matriz de GLCM para una img, se concatenan y son proporcionadas al clasificador de máquina de vectores de

soporte no lineal para determinar si es cancerígeno o no ya que los nódulos no se pueden separar linealmente, además el SVM es entrenado y probado con validación cruzada de 10 veces para la verificación del error en el rendimiento del modelo propuesto.

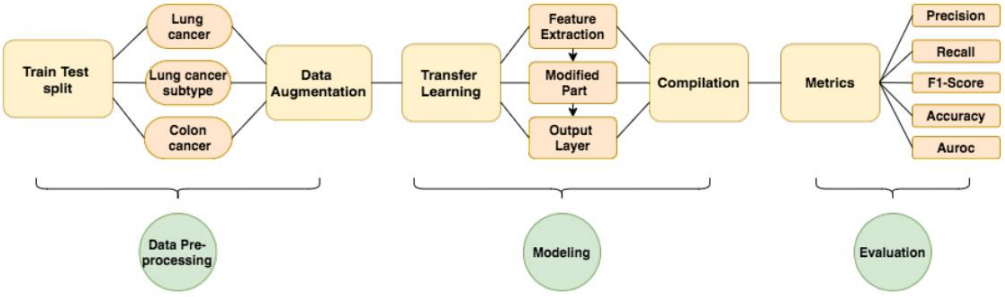


Figura 43. Diagrama de trabajo de [33]

En [33] se puede apreciar que se divide en 3 etapas preprocesamiento de datos, modelamiento y evaluación, primeramente denomina prueba de tren dividida se especifica en conjunto de datos usado LC25000 que contiene imágenes pulmonares (15000 instancias) y de colonoscopia (10000 instancias), se usa la partición 80 – 20 para las clasificación binarias son 3 la primera es para identificar el cáncer de pulmón, la segunda encontrar subtipos del cáncer, la tercera para reconocer el cáncer de colon, luego en el aumento de datos las img de la base de datos se recortan a un tamaño de 224x224, también se utiliza una tubería para el aumento de la img mediante la librería imgaug. En la etapa 2 la transferencia de aprendizaje se transfiere los pesos de un modelo existente a otro para reducir el consumo computacional y aumentar el rendimiento del sistema aplicado a tres actividades de clasificación binaria estas se dividen en extracción de características, modificación y salida, en la primera se recopilan las características más relevantes por medio del ajuste de modelos a través de 8 arquitectura CNN entrenadas previamente, en la segunda se agrega una capa enlazada y una plana para generar un vector de entidades, en la tercera se agrega a la capa de salida la capa mencionada previamente. En la última sub etapa del modelamiento se realiza la compilación puesto que los algoritmos deben de optimizarse y se eligió Adam debido a que es el que mejor funciona en modelo a aprendizaje profundo. Finalmente, en la última etapa se medirá el modelo teniendo en cuenta los indicadores mostrados en la figura 43.

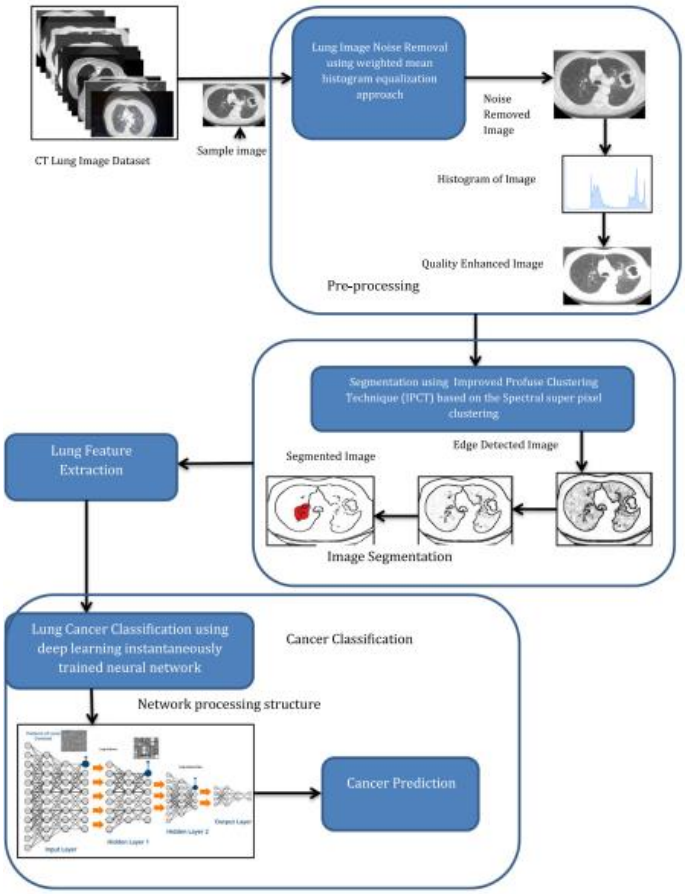


Figura 44. Diagrama de trabajo de [67]

En la figura 44 se detalla el proceso de trabajo de la investigación [67] en la que se recopila información, seguidamente de preprocesamiento de datos, luego segmentación de imagen y concluye con la etapa de clasificación. Inicialmente se adquiere un conjunto de datos CIA de imágenes (5043 instancias) en formato Dicom y se divido para entrenamiento y prueba 3000 y 2043 correspondientemente. En el procesamiento se examinan las instancias en términos de predicción de ruido, contraste y técnicas de histograma para mejorar la calidad. La segmentación de imagen medica TC se realiza por medio de una técnica de agrupación difusa mejorada IPCT de manera concreta el algoritmo funciona acorde con el grafico no definido de cada pixel prediciendo eficazmente los bordes de la imagen, se extraen atributos estadísticos y espectrales de la región para identificar el cáncer de forma correcta. En la sección de características espectrales la parte segmentada anteriormente se pasa a esta etapa donde se extraen atributos como asimetría, media, desviación estándar y curtosis debido a que detecta eficazmente las características del cáncer de pulmón, luego son procesadas mediante una red neuronal entrenada

instantáneamente de aprendizaje profundo DITNN. Finalmente, en la fase de clasificación se usa DITNN, durante el entrenamiento se usa las TC o img segmentada para identificar los bordes y características relevantes, generalmente la red contiene 150 capas ocultas.

¿Cuáles son las métricas que se utilizan para medir los modelos?

Dependiendo de cada investigación se muestra un detalle especifico, formula, tabla de resultas, comparación entre algoritmos y promedio de las mismas. En estudio [61] utilizan redes neuronales convolucionales y consideran sensibilidad que representa la correlación entre la imagen medica que posee nódulo pulmonar y la clasificación correcta con nódulo, especificidad de manera similar al anterior pero esta considera la correlación y clasificaron con aquellas que no poseen nódulo, precisión que es el correlación y clasificación de aciertos correctamente y F1 score que representa una forma de medición entre la presión y sensibilidad de manera similar en [18] se opta los tres primeros añadiendo a su estudio la exactitud que representa la cantidad de aciertos correctos, en [36] eligen a los tres anteriores sumándole la AUC que es la medida para evaluar el desempeño de los límites, aunque en los estudios mencionados no se especifica el detalle de los mismos como en la investigación [16] se realiza un detalle sobre las métricas que usan y como se determina el valor de cada una especificando la fórmula de exactitud, precisión y sensibilidad algo que usualmente solo se puntualiza en libros, comparándola con los artículos [65] [49] [34] estos complementan ampliamente los resultados indicando el significado de los valores de cada expresión, definición de las métricas, detalle en tablas y gráficos entre los ellos. Aquellas mencionadas principalmente son las más habituales no obstante la investigación [53] considero Puntuación media de FROC que es la media de la sensibilidad entre la cantidad de falsos positivos por escaneo (FPs/Scan) y el tiempo de ejecución medido en minutos para destinarlo a una comparación entre sistemas asistidos del cáncer de pulmón.

¿Cuáles son los trabajos de investigación que han obtenido mejores resultados en la clasificación de cáncer de pulmón utilizando imágenes tomográficas?

En la etapa de calificación de las investigaciones usualmente no se superaba el porcentaje indicado respecto a las métricas de evaluación y otras solo

contemplan un indicador como en [13] el porcentaje promedio de precisión es de 95% siendo un porcentaje aceptable del mismo modo [10] obtiene un 97,4% en el mismo indicador pero no se consideran otras métricas que ofrecen una visión más clara en la evaluación, en el caso de [65] si se muestran tres métricas de manera clara: precisión, especificidad y sensibilidad alcanzando un porcentaje entre 93% a 95%. La técnica propuesta en [14] obtuvo resultamos análogos al antes mencionado en los mismos términos 93%, 95 % y 86% en la máquina de vectores de soporte (SVM) y para K vecinos más próximo (KNN) 91%, 93% y 82% respetivamente. En el estudio de [7] emplearon como clasificador de estadio del cáncer de pulmón a KNN obtenido un valor de 98,15% de presión, además de ello consideran características de la imagen: área, perímetro, centroide y etapa, una limitante es que solo pueden clasificar el cáncer en el segundo estadio por lo que proponen recopilar mayor cantidad de datos para clasificar las diferentes etapas del cáncer de pulmón.

3.3.2. Adquirir el data set de cáncer de pulmón

En este punto de la investigación se documenta la adquisición del dataset puesto que mediante ello se podrán realizar el preprocesamiento de datos, selección de técnicas, implementación y pruebas. Se realizó a partir de los artículos científicos mencionados con anterioridad en ellos se emplean varios conjuntos de datos como Luna-16, LIDC, TCGA-LUAD, LIDC-IRDI, Lung Cancer Data Set del repositorio UCI, the Kaggle Data Science Bowl, SPIE-AAPM Lung CT Challenge dataset los cuales se han detallado en la tabla 13, además de ellos existe una colección de datos en https://www.cancerimagingarchive.net/collections/ en la que podemos encontrar el nombre del dataset, el tipo de cáncer, número de ejemplares, tipo de img y si el acceso del conjunto de datos está protegido o público.

Durante la revisión de los artículos la mayoría emplea el conjunto "LIDC-IRDI" para el cáncer de pulmón que consta de 1010 participantes, 1308 instancias, situada en la zona del pecho, las img que contienen son tomografía computarizada CT, radiografía computarizada CR y rayos x digitales DX no obstante estos no están etiquetados, ni separados por carpetas para clasificar de forma correcta.

Dentro de la página se puede apreciar los mencionado anteriormente algunas de las bases de datos y otras características como la especie que en este caso es "Humanos", nombre de la entidad que proporciono la información, el estado y la fecha de actualización.

Tipo de datos	Descargar todo o Consultar/Filtrar
Imágenes (DICOM, 125GB)	Download Search
Compendio de metadatos DICOM (CSV)	Download
Anotaciones/segmentaciones del radiólogo (formato XML) (Nota: consulte pylidc para obtener ayuda con el uso de estos datos)	Download
Lista de tamaños de nódulos (web)	Search
Recuento de nódulos por paciente (XLS)	Download
Diagnósticos del paciente (XLS)	Download

Figura 45. Descripción del conjunto de datos LIDC-IRDI

Después de seleccionar el conjunto de datos, podemos apreciar el apartado de acceso a datos y una descripción del archivo csv, en esta investigación se trabajará

con el conjunto de datos “Chest CT-Scan images Dataset”

Realizando una comparación entre los dataset disponibles y las características de cada uno “LIDC-IDRI” posee una mayor cantidad de imágenes médicas y número de pacientes, pero el Conjunto de datos seleccionado posee un orden, etiquetando cada tipo de cáncer y separando el conjunto de entrenamiento, prueba y validación.

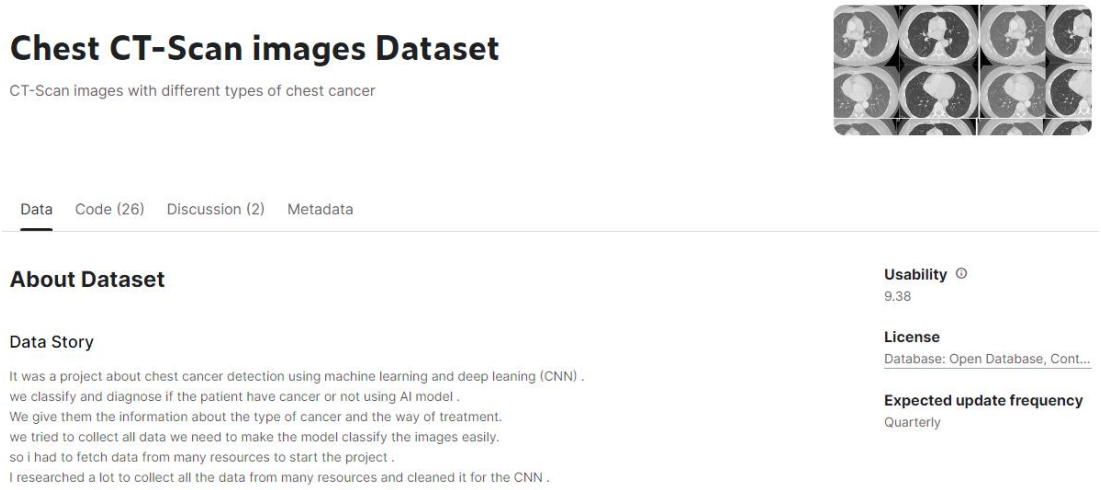


Figura 46. Conjunto seleccionado

Tabla 14. Conjuntos de datos de selección.

Conjunto de datos	**Tipo de img**	**Participantes**	**Numero de img**	**Tamaño de img GB**
SPIE-AAPM Lung CT Challenge	Tomografía computarizada CT	70	22,489	12.1
TCGA-LUAD	Tomografía computarizada CT, Terapia física PT y Medicina Nuclear NM	69	48,931	18.3
LIDC-IDRI	Tomografía computarizada CT, radiografía computarizada CR	1010	244,527	124

	y Rayos x digitales DX		
APOLLO-5-LUAD	Tomografía computarizada CT	124	3.3
Chest CT-Scan images Dataset	Tomografía computarizada CT	1,000	0.124

Para descargar el conjunto seleccionado ingresamos a la página de kaggle por medio del siguiente enlace https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images y en el apartado de API creamos una nueva ya que por medio de ella podremos descargarla en nuestro entorno de trabajo, este archivo contiene nuestro nombre de usuario y clave.
En las siguientes líneas de código se detalla el funcionamiento de cada una, en la primera importamos una librería para trabajar con archivos después de seleccionar el archivo, nos permite poder realizar la descarga con los comandos de kaggle se descarga en un archivo chest-ctscan-images.zip posteriormente se descomprime en una carpeta Data que contiene el conjunto de prueba, validación y entrenamiento cada subcarpeta posee los tipos de cáncer (adenocarcinoma, normal, célula larga y escamosa) adicionalmente si desplegamos una subcarpeta como Adenocarcinoma podemos seleccionar las img y visualizarla.

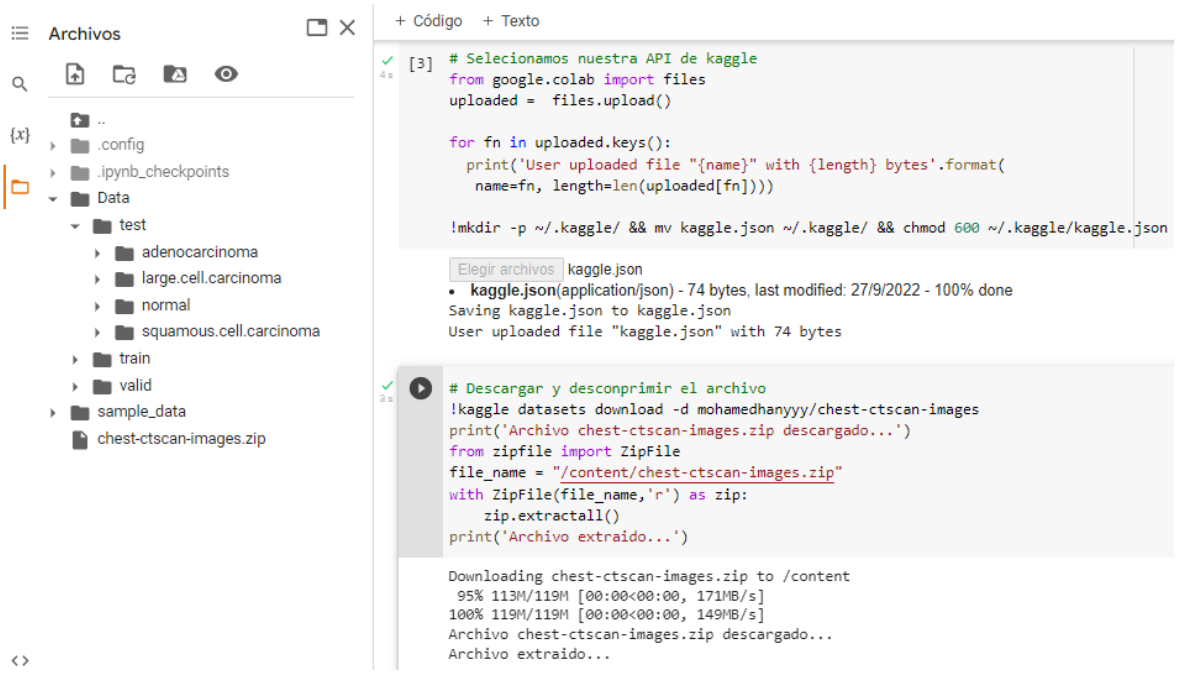


Figura 47. Obtener el conjunto de datos

Figura 48. Visualizar img de las carpetas

Ahora para poder ver la distribución de los datos mediante gráficos primero obtenemos el tamaño de cada subcarpeta, lo asignamos a su etiqueta correspondiente y mostramos el grafico.

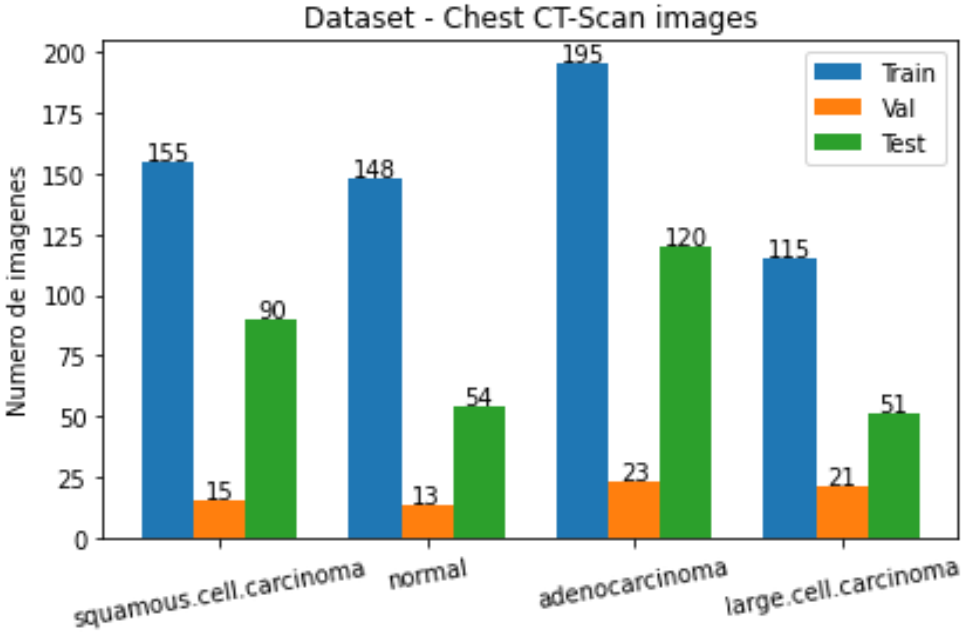


Figura 49. Distribución del conjunto de datos

3.3.3. Seleccionar las técnicas de Aprendizaje automático para detectar el cáncer de pulmón

Por medio de ImageNet se han propuesto varias arquitecturas en los últimos años la siguiente figura representa un análisis comparativo entre arquitecturas de redes neuronales profundas en el eje Y representa la precisión de cada una en el eje X el nivel de complejidad.

Podemos apreciar la basta cantidad de modelos profundos desarrollados por lo que se selecciona aquellos que poseen mejores resultados y los más empleados actualmente en la clasificación de imágenes: Mobilenetv2, Densenet201, Efficientb4, Resnet50.

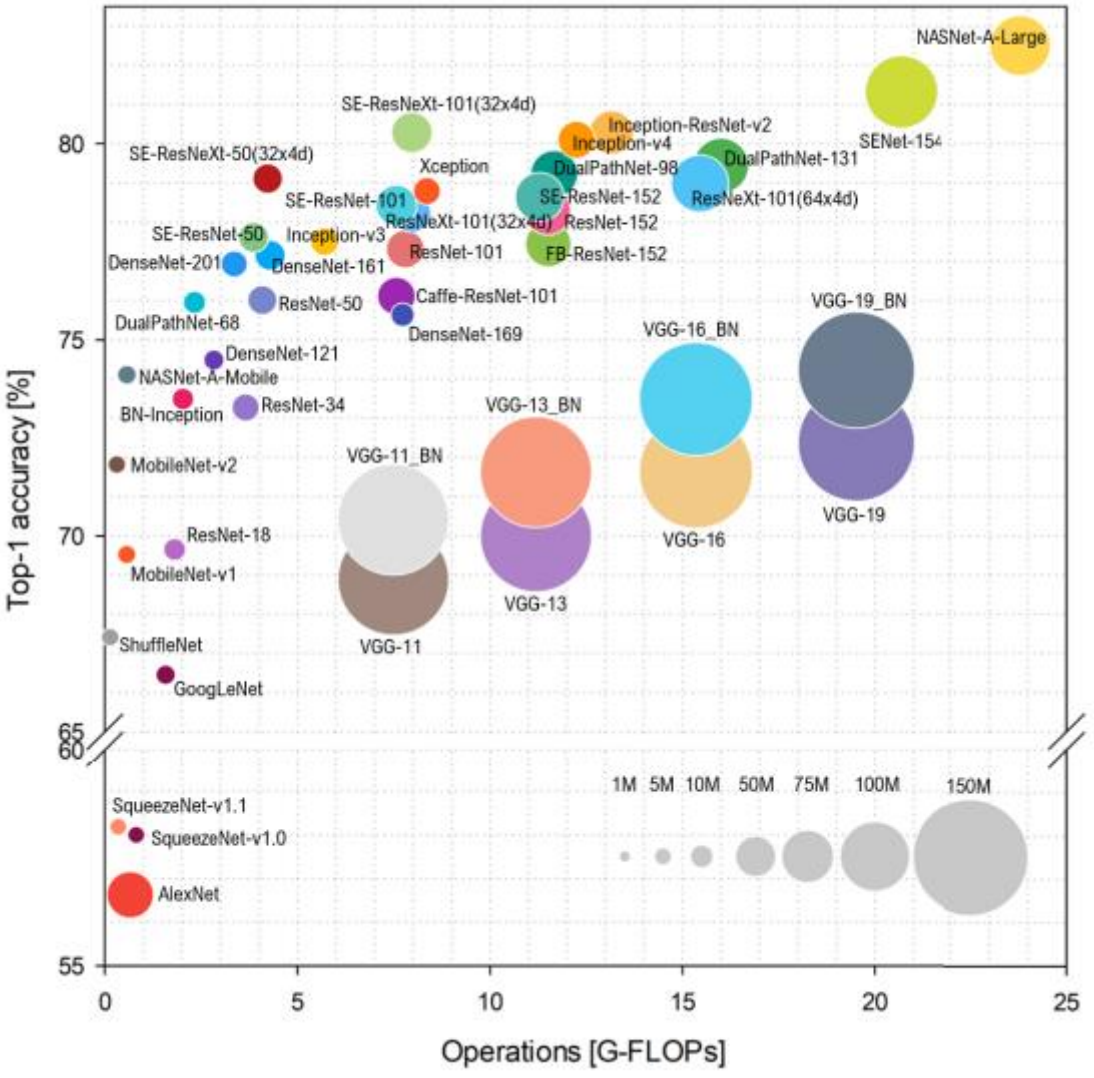


Figura 50. Comparación de arquitecturas de aprendizaje profundo.

3.3.4. Implementar las técnicas de aprendizaje automático para detectar el cáncer de pulmón

En esta sección seleccionamos como entorno de trabajo Google Colab y lenguaje de programación Python en el que podemos usar una máquina virtual y configurar el entorno de ejecución a GPU para ejecutar el código de forma acelerada.

Antes de pasar el conjunto de datos tenemos que importar algunas librerías para poder usar las arquitecturas, configurarlas, llamar a las métricas de evaluación, procesar las img, compilar el modelo, establecer un punto de control, parada temprana, entrenar y validar el modelo, en las siguientes líneas se importan las librerías necesarias para la implementación

```
# Keras: Librería de Redes Neuronales de Código Abierto
import tensorflow.keras as keras
# ImageDataGenerator: Generar lotes de datos de imágenes tensoriales con aumento de datos en tiempo real.
# load_img: Carga una imagen en formato PIL.
# img_to_array: Convierte una instancia de imagen PIL en una matriz Numpy.
from tensorflow.keras.preprocessing.image import ImageDataGenerator, load_img, img_to_array
# Sequential: agrupa una pila lineal de capas en un tf.keras.Model.
from tensorflow.keras.models import Sequential
from keras.utils import np_utils
# Dense: Sólo una capa NN normal densamente conectada.
# Activation: Aplica una función de activación a una salida.
# Flatten: Aplana la entrada. No afecta al tamaño del lote.
# Dropout: Apagar aleatorialmente neuronas
# BatchNormalization: Capa que normaliza sus entradas.
# Conv2D: Capa de convolución 2D
# MaxPooling2D: Operación de pooling máximo para datos espaciales 2D.
from tensorflow.keras.layers import Dense, Activation, Flatten, Dropout, BatchNormalization, Conv2D, MaxPooling2D
# ModelCheckpoint: Llamada de retorno para guardar el modelo Keras o los pesos del modelo con cierta frecuencia.
# EarlyStopping: Detenga el entrenamiento en caso alguna métrica monitoreada deje de optimizar.
```

```
from tensorflow.keras.callbacks import ModelCheckpoint,
EarlyStopping
# regularizers: Clase base del regularizador.
# optimizers: Clases optimizadoras incorporadas.
from tensorflow.keras import regularizers, optimizers
# libreria de IA desarrollada por google
import tensorflow as tf
# Instanciar la arquitectura de cada modelo
from tensorflow.keras.applications import ResNet50, DenseNet201,
EfficientNetB4, MobileNetV2
# Modelos de Keras
from tensorflow.keras.applications import resnet, densenet,
efficientnet, mobilenet_v2
# Para visualizar gráficos de los modelos
from tensorflow.keras.callbacks import TensorBoard
# El modelo permite agrupar las capas en un objeto con
características de entrenamiento.
from tensorflow.keras import Model
# Para la generación de gráficos y trabajar con arreglos numéricos
import matplotlib.pyplot as plt
import numpy as np
# Para manipulación y análisis de datos
import PIL
import os
import pandas as pd
import cv2
```

Esta figura representa aquellas librerías que se usan Keras nos permite hacer uso de las redes neuronales, ImageDataGenerator permite realizar transformaciones sobre las img, Sequential para agregar capas, Dense agrega capas normales, Activation representa a las funciones de activación existentes, Flatten aplana la entrada a la red, Dropout apaga de forma aletoria algunas redes, BatchNormalization normaliza las entradas, Conv2d para agregar capas de convolución, ModelCheckPoint guarda los pesos del modelo, EarlyStopping detendrá el entrenamiento cuando alguna métrica deje de mejorar, luego importamos las arquitecturas y TensorBoard para ver las

métricas del modelo mediante graficos, también agregamos otras librerías adicionales para las img, la manipulación y análisis de datos.

Después agregamos una extensión de Python para que nos indique el tiempo que demora cada ejecución puesto que después de 40min o más el tiempo de ejecución que muestra Colab se reinicia o no se muestra.

```
!pip install ipython-autotime
%load_ext autotime
```

Pasamos a la configuración de cada arquitectura como el tamaño de la img que recibe, los canales de colores, el número de clases que para este caso son los 4 tipos de cáncer, es opcional usar el aumento de datos para los modelos ya que las TC siempre se entregan de la misma forma, pero experimentar con girar, recortar, acerca, alejar o mover a los lados y agregarle una función de preprocesamiento podría generar mejores resultados si comparamos las métricas de los modelos con aumento o sin ello.

```
#Definicion de parámetros, numero de clases, tamaño de #imagen y de lote
image_shape = (460,460,3)
N_CLASSES = 4
BATCH_SIZE = 32

#Nota deberá de seleccionar si usar o no el Aumento de #datos, Arquitectura_X simboliza las arquitecturas #definidas en la muestra de la investigación para realizar #sus pruebas deberán de modificar este nombre

#Entrenamiento sin Aumento de datos, para la función de #procesamiento se tendrá que llamar a los modelos y a su #función por ejemplo: resnet.preprocess_input, #efficient.preprocess_input, mobilenet.preprocess_input y #densenet.preprocess_input

train_datagen = ImageDataGenerator(dtype='float32', preprocessing_function=Arquitectura_X.preprocess_input)
```

```
        #Entrenamiento con Aumento de datos, en la que rotamos, #movemos
hacia los costados, invertimos, acercamos o #alejamos la img
        train_datagen = ImageDataGenerator(
        dtype='float32',
        rotation_range=10,
        width_shift_range=0.1,
        height_shift_range=0.1,
        shear_range=0.1,
        zoom_range=0.1,
        horizontal_flip=True,
        fill_mode='nearest'
        ,
        #Para definir la función de preprocesamiento tenemos que #llamar al
modelo e indicar esa función
        preprocessing_function= Arquitectura_X.preprocess_input
        )

        #Ahora almacenamos esos conjuntos generados y la asignamos #a
unas variables para usarlas en el entrenamiento, prueba #y validación, del
conjunto generado se aplicará o no el #aumento de datos y preprocesamiento,
después establecemos #los parámetros definidos inicialmente

        #Entrenamiento
        train_datagen.flow_from_directory(train_path,
batch_size = BATCH_SIZE,                    target_size =
(460,460),                    class_mode = 'categorical')

        #Validacion
        valid_datagen = ImageDataGenerator(dtype='float32',
preprocessing_function= Arquitectura_X.preprocess_input)
        valid_generator = valid_datagen.flow_from_directory(valid_path,
batch_size = BATCH_SIZE,                    target_size =
(460,460),                    class_mode = 'categorical')

        #Prueba
        test_datagen = ImageDataGenerator(dtype='float32',
```

```
preprocessing_function=efficientnet.preprocess_input)
        test_generator = test_datagen.flow_from_directory(test_path,
batch_size = BATCH_SIZE,                    target_size =
(460,460),                    class_mode = 'categorical')
```

Las configuraciones presentadas son similares para cada una de las arquitecturas Resnet50, Efficientnetb4, Densenet201 y Mobilenetv2

Una vez realizado esto de forma similar para cada arquitectura se procede a instanciar el modelo definimos si incluimos la última capa, la agrupación para reducir la muestra de entrada, inicializar los pesos con ImageNet o no, definir que capas vamos a usar y si las entrenamos, además podemos agregar a ello otras capas, funciones de activación, normalización, apagar aleatoriamente las neuronas, aplastar la entrada, entre otras.

#Para instanciar las arquitecturas llamamos al modelo y #establecemos la configuración, incluir la última capa, #capa de agrupación máxima o promedio, mantener los pesos #de imagenet, la entrada de la img y definir que capas son #entrenadas o no

```
        #ResNet50
        res_model = ResNet50(include_top=False, pooling='avg',
weights='imagenet', input_shape = (image_shape))
        for layer in res_model.layers:
            if 'conv5' not in layer.name:
              layer.trainable = False

        #EfficientNetB4
        efficient_model = EfficientNetB4(include_top=False, pooling='avg',
weights='imagenet', input_shape = (image_shape))
        for layer in efficient_model.layers:
            if 'block7'not in layer.name and'top' not in layer.name:
                layer.trainable = False

        #DenseNet201
        dense_model = DenseNet201(include_top=False, pooling='avg',
```

```
weights='imagenet', input_shape = (image_shape))
        for layer in dense_model.layers:
            if 'conv5' not in layer.name:
                layer.trainable = False

        #MobileNetV2
        mobilenetv2 = MobileNetV2(include_top=False, pooling='avg',
weights='imagenet', input_shape = image_shape )
        # mobilenetv2.summary()

        for layer in mobilenetv2.layers:
            if 'block_15'not in layer.name and'block_16' not in layer.name:
                layer.trainable = False
```

Definimos las métricas de evaluación en un arreglo para posteriormente usarlas en el entrenamiento, estas métricas se usarán para todos los modelos.

```
#Metricas de evaluación de modelos, f1_score es importada #del backend de
keras ya que actualmente no es posible #llamarla como a las otras

        import keras.backend as KE
        def f1_score(y_verdadero, y_prediccion):
            verdaderos_positivos = KE.sum(KE.round(KE.clip(y_verdadero *
y_prediccion, 0, 1)))
            posibles_positivos = KE.sum(KE.round(KE.clip(y_verdadero, 0, 1)))
            predicciones_positivas = KE.sum(KE.round(KE.clip(y_prediccion, 0,
1)))
            precision = verdaderos_positivos / (predicciones_positivas +
KE.epsilon())
            sensibilidad = verdaderos_positivos / (posibles_positivos +
KE.epsilon())
            f1_valor =
2*(precision*sensibilidad)/(precision+sensibilidad+KE.epsilon())
            return f1_valor

        # Definimos las métricas de evaluación para usarlas en el
```

entrenamiento

```
metrics_eval = [
    tf.keras.metrics.BinaryAccuracy(name='accuracy'),
    tf.keras.metrics.Precision(name='precision'),
    tf.keras.metrics.Recall(name='recall'),
    f1_score,
]
```

Definimos el optimizador, la tasa de entrenamiento, función de perdida (loss) y las métricas (metrics) para compilar el modelo, establecemos un punto de control que guardará el mejor modelo durante la ejecución y este se podrá usar posteriormente para realizar predicciones, la parada temprana monitorea la ejecución para evitar que el modelo continue entrenándose cuando ha dejado de mejorar las métricas

```
# ModelCheckpoint: Devolución de llamada para guardar el modelo Keras o los pesos del modelo con alguna frecuencia.
# filepath: ruta para guardar el archivo del modelo
# monitor: El nombre de la métrica a monitorear
# save_best_only: Si, solo se guarda cuando el modelo se #considera el "mejor" y no se sobrescribirá el mejor modelo más reciente de acuerdo con la cantidad monitoreada.
#Nota: LC_NombreArquitectura_Descripcion debe de ser modificado por el nombre del modelo, su descripción (Con/Sin aumento de datos) y extensión (hdf5 o h5), por ejemplo:
LC_Resnet50, LC_Efficientnetb4, LC_Densenet201, LC_Mobilenetv2, LC_Resnet50_DA, LC_Efficientnetb4_DA, LC_Densenet201_DA y LC_Mobilenetv2_DA

checkpointer = ModelCheckpoint(filepath='./LC_NombreArquitectura_Descripcion.h5', monitor='val_loss', verbose = 1, save_best_only=True)

# EarlyStopping: Deje de entrenar cuando una métrica monitoreada haya dejado de mejorar.
# Patience: Representa el N° de épocas después de las cuales se
```

detendrá el entrenamiento en caso no mejore.

verbose (0/1): Modo de verbosidad, 0 o 1. El modo 0 es silencioso y el modo 1 muestra mensajes cuando la devolución de llamada realiza una acción.

```
early_stopping = EarlyStopping(verbose=1, patience=15)
```

Estos se guardarán por defecto en una carpeta llamada “log” para usar tensorboard y ver cómo va el entrenando el modelo de forma gráfica (visualizar métricas)

En el entrenamiento le enviamos las configuraciones previas para y por consola veremos cómo va aprendiendo según se vean las métricas

```
#La variable de tensorboard se envia en el arreglo de "callbacks" (hay otros tipos de callbacks soportados)
#En este caso guarda datos en la carpeta indicada en cada epoca, de manera que despues
#Tensorboard los lee para hacer graficas
tb_modelo_x = TensorBoard(log_dir='logs/TensorBoard_modelo_x ')

history_modelo_x = model.fit(train_generator,
                steps_per_epoch = 20,
                epochs = 200,
                verbose = 1,
                validation_data = valid_generator,
                callbacks = [checkpointer, early_stopping, tb_ modelo_x])
```

3.3.5. Realizar pruebas sobre las técnicas implementadas.

Después del entramiento entramos en la validación de los modelos, asi mismo ejecutamos o recargamos tensorboard para visualizar las métricas de nuestros modelos

```
# Las validaciones de Resnet50, EfficientB4, Densenet201 y # MobileNetv2 se realizan
Result_x = model_x.evaluate(test_generator)
```

Después para realizar otras pruebas o tenemos que realizarlos de forma similar al procedimiento explicado, esto es alojado en TensorBoard de ello generaremos graficas.

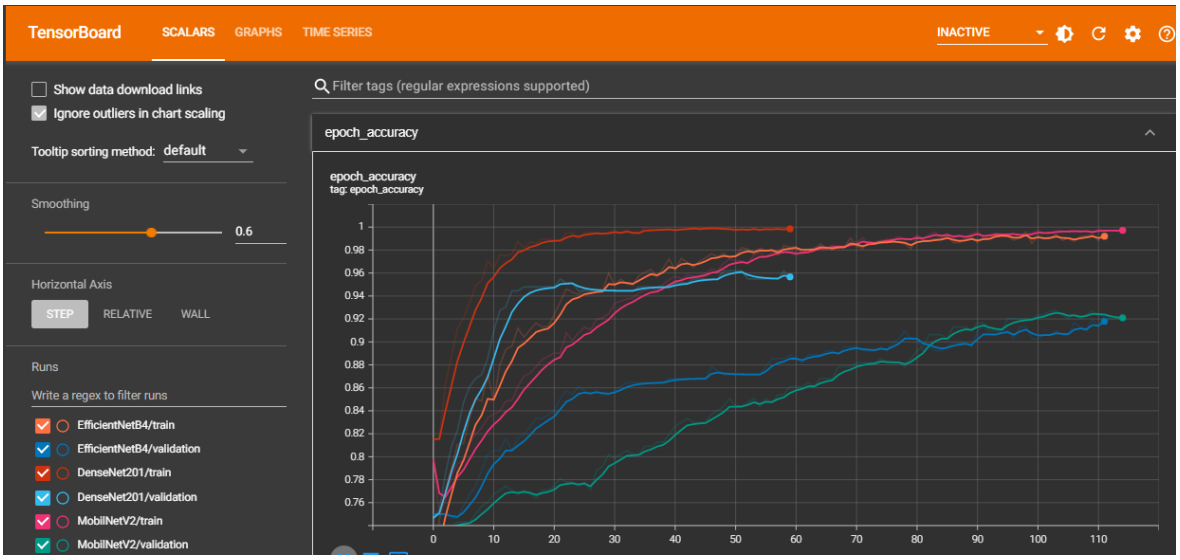


Figura 51. Vista de comparación entre modelos con TensorBoard

3.3.6. Desarrollar una aplicación web

Para realizar el desarrollo de la aplicación usamos como metodología XP y framework Streamlit que permite crear y compartir aplicaciones de forma rápida, dentro de su documentación se encontrara información para poder agregar lo que necesitemos, en nuestro entorno de trabajo instalamos streamlit, para posteriormente escribir en un archivo con extensión py "AppLungCancer.py" importa librerías de tensorflow y del mismo framework luego definimos una función para cargar el modelo con mejores resultados "Efficientb4_DA", escribimos un título, colocamos una entrada de archivos para formato de img.

Una vez seleccionada la img será enviada a la función de predicción la cual recibe la img, definimos el tamaño con la que fue entrenada (460x460), después el modelo preprocesa la img (función de preprocesamiento), pasando a la predicción y retornando la etiqueta o el tipo de cáncer para ser mostrara en la aplicación, puede visualizar el cronograma de desarrollo en el Anexo 5.

Figura 52. Desplegar modelo

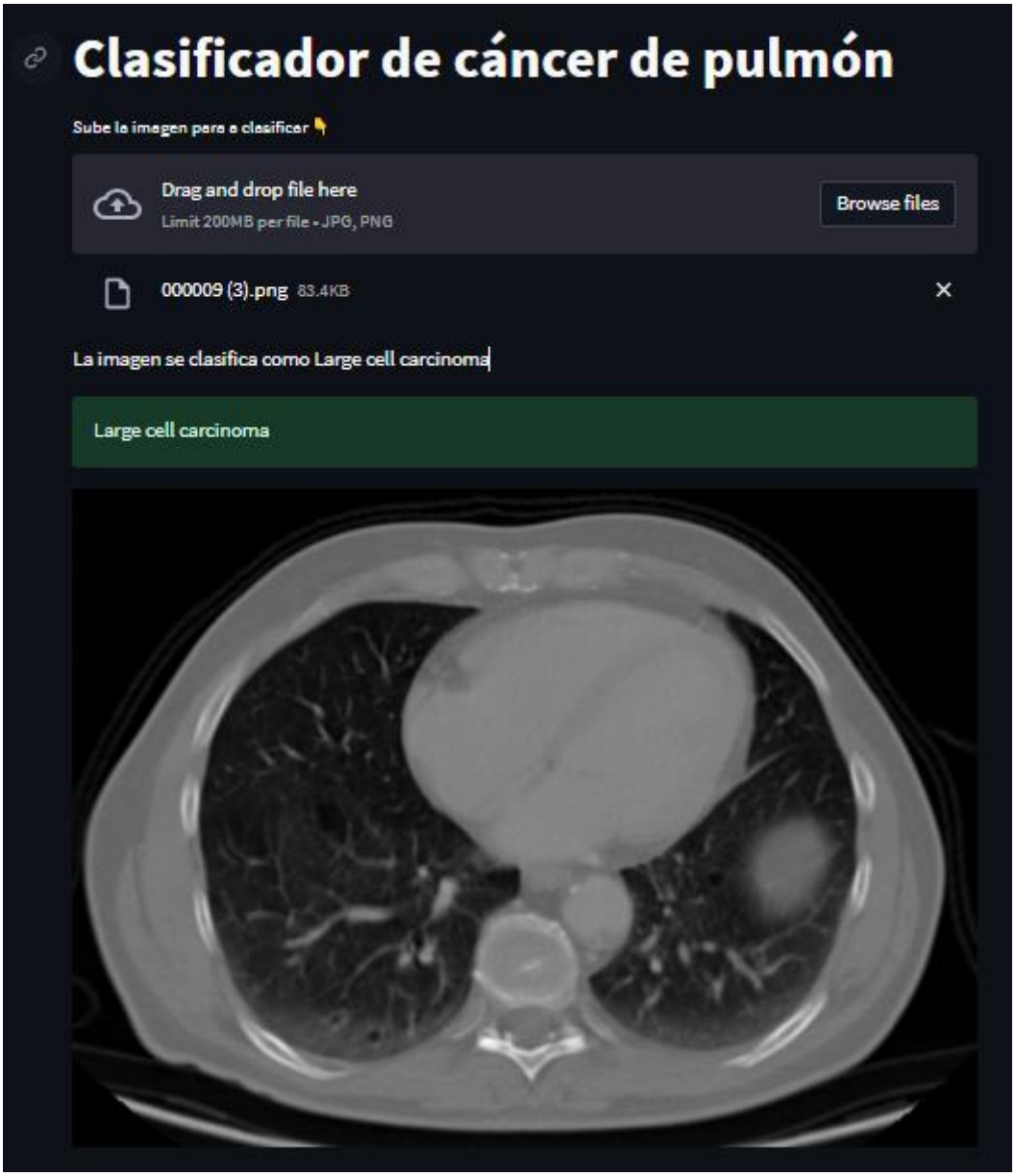


Figura 53. Predicción del sistema

3.3.6.1. Planificación

Historias de usuario: Los puntos estimados se consideran desde 1 a 5 en un nivel de dificultad, el número de iteraciones considerados para la etapa de iteraciones (1), el riesgo durante el desarrollo y nivel de prioridad se consideran para ambos altos (A), medio (M) y bajo (B).

Tabla 15. Historia de Usuario

Numero: 1	**Nombre:** Gestión de Clasificador de Cáncer
Usuario: Administrador/Usuario	
Prioridad de negocio: A	**Puntos estimados:** 4
Riesgo en desarrollo: A	**Iteración:** 1
Descripción: Se deberá de poder subir las imágenes médicas al módulo de clasificación posteriormente se mostrará el resultado de la imagen	
Observaciones:	

Valores: Los valores que considera esta metodología son cinco el primero “Simplicidad” hace referencia que se hará lo necesario y lo que se solicite, no más, el segundo “Comunicación” significa que los involucrados trabajaran en conjunto en los requerimientos, desarrollo del código y solución de problemas, “Comentarios” durante las entregas establecidas de avances del proyecto se tendrá en cuenta las opiniones de los usuarios y se realizara cambios requeridos, “Respeto” se considera valioso el aporte de todas las partes, finalmente el “Coraje” representa decirla verdad sobre el cronograma y adaptación a cambios cuando ocurran.

Criterios de adaptación

Tabla 16. Criterios de Adaptación

Requerimientos	Criterios de aceptación
Gestión de Clasificador de Cáncer	Acceso al módulo y permite subir la imagen para realizar la clasificación

Plan de iteración: Las iteraciones permiten que cada historia de usuario se complemente a medida que se desarrolle y analice cada uno siendo una medida clara de avance, la primera iteración es aquellas que son necesarias para el posterior funcionamiento de posibles iteraciones en un futuro desarrollo.

Es importante aclara que el desarrollo de estas iteraciones se considera desde la etapa de diseño y codificación entre las fechas 06/12/22 – 15/12/22 (2 semanas)

Tabla 17. Plan de Iteración

Iteración	Historia de Usuario	Actividades	Puntos estimados	Tiempo Estimado
1ra	Gestión de Clasificador de Cáncer	Diseño de interfaz Programación Modificaciones	4	2 semanas

3.3.6.2. Diseño

Diseño Simple: Durante el proceso de diseño de los prototipos, se dará prioridad a la simplicidad y funcionalidad del sistema, buscando representar estas características en el modelo de la aplicación. Se buscará desarrollar una interfaz que sea fácil de entender y utilizar para los usuarios, optimizando la experiencia de usuario y garantizando un flujo de trabajo eficiente. La simplicidad en el diseño de los prototipos permitirá una implementación más rápida y una mejor comprensión de las funcionalidades de la aplicación, lo que a su vez facilitará el proceso de desarrollo y reducirá la posibilidad de errores. Al plasmar estas consideraciones en el modelo de

la aplicación, se espera lograr una solución intuitiva y accesible para los usuarios finales, mejorando su satisfacción y aumentando la usabilidad del sistema.

Tarjeta CRC

Tabla 18. Tarjeta CRC

Administrador / Usuario	
Descripción: Se describe las acciones a realizar por el usuario en relación al sistema	
Responsabilidades	
Nombre	**Colaborador**
Subir imagen, visualizar resultados	

Prototipos

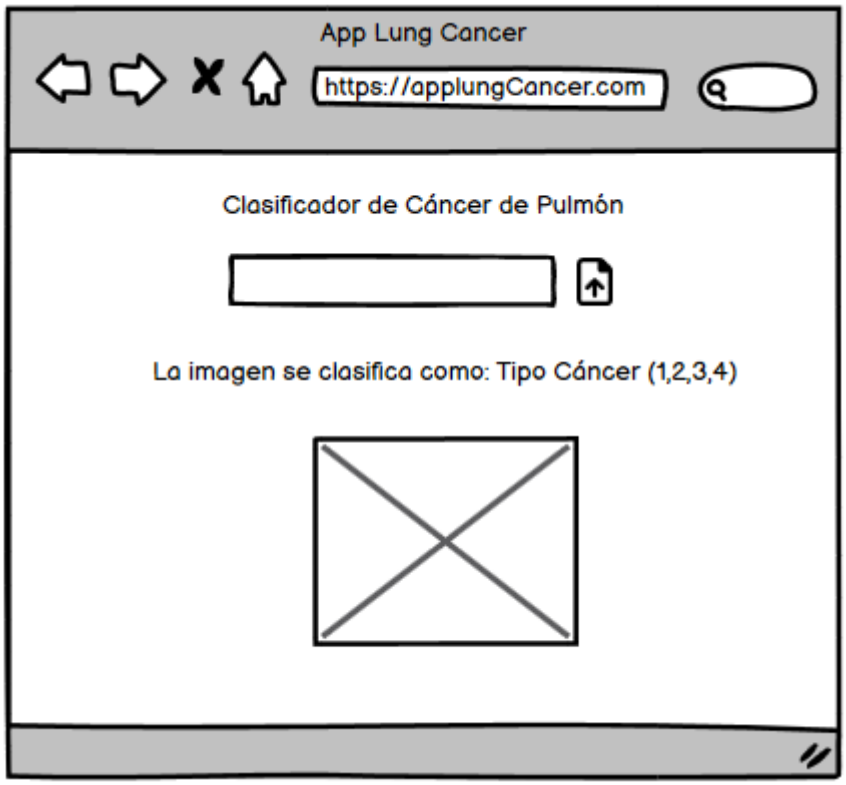


Figura 54. Prototipo Inicial

3.3.6.3. Codificación

Programación y rediseño: Para este punto ya tenemos nuestro modelo EfficientNet_DA entrenado y hemos guardado el modelo en un archivo con extensión .h5 para usarlo temporalmente con streamlit, El código proporcionado es un clasificador de cáncer de pulmón implementado utilizando la biblioteca Streamlit y el modelo EfficientNet de TensorFlow. A continuación, se presenta un resumen de lo que hace el código:

- Importación de bibliotecas: Se importan las bibliotecas necesarias, como Streamlit, TensorFlow, y otras bibliotecas relacionadas con el procesamiento de imágenes.
- Carga del modelo: La función load_model() carga el modelo pre-entrenado de EfficientNet para clasificar imágenes de cáncer de pulmón. El modelo se carga desde un archivo HDF5 / H5.
- Interfaz de usuario: Se crea la interfaz de usuario utilizando Streamlit. Se muestra un encabezado con el título "Clasificador de cáncer de pulmón".

- Carga de la imagen: Se proporciona un botón para subir una imagen para clasificar. El archivo de imagen subido se lee utilizando file_uploader() de Streamlit.
- Preprocesamiento de la imagen: La imagen subida se redimensiona y se convierte en un arreglo NumPy. Se realiza el preprocesamiento necesario en la imagen, incluida la normalización.
- Clasificación de la imagen: Se utiliza el modelo cargado para realizar la clasificación de la imagen. El modelo devuelve la clase predicha, que representa el tipo de cáncer de pulmón.
- Resultados: Se muestra la clase predicha junto con un mensaje indicando cómo se clasifica la imagen. Además, se muestra la imagen subida junto con el resultado de la clasificación.

En resumen, el código implementa un clasificador de cáncer de pulmón utilizando el modelo EfficientNet. Permite a los usuarios cargar una imagen y obtener la clasificación del tipo de cáncer de pulmón asociado a la imagen.

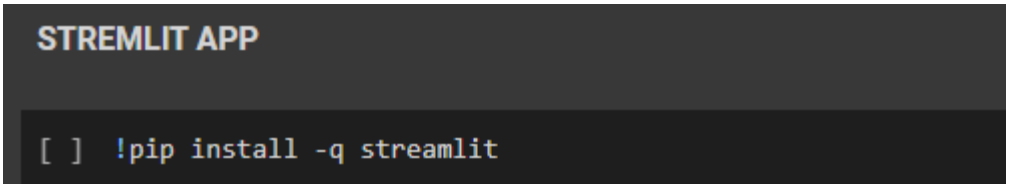


Figura 55. Fragmento de Código Inicial

```
%%writefile AppLungCancer.py
import streamlit as st
import tensorflow as tf
from tensorflow.keras.applications.imagenet_utils import preprocess_input, decode_predictions
from tensorflow.keras.preprocessing import image
import matplotlib.pyplot as plt

from tensorflow.keras.applications import resnet, vgg16 , vgg19, densenet, efficientnet, mobilenet_v2

@st.cache(allow_output_mutation=True)
def load_model():
  #model=tf.keras.models.load_model('/content/chest_CT_SCAN-ResNet5099.h5')
  #chest_CT_SCAN-MobilNetV2
  model=tf.keras.models.load_model('/content/chest_CT_SCAN-MobilNetV2.hdf5')
  return model
with st.spinner('Model is being loaded..'):
  model=load_model()

st.write("""
         # Clasificador de cáncer de pulmón
         """
         )
```

Figura 56. Fragmento de Código Principal – Parte 1

```
file = st.file_uploader("Sube la imagen para a clasificar \U0001F447", type=["jpg", "png"])
import cv2
from PIL import Image, ImageOps
import numpy as np

st.set_option('deprecation.showfileUploaderEncoding', False)
def upload_predict(upload_image, model):

        size = (460,460)
        image = ImageOps.fit(upload_image, size, Image.ANTIALIAS)
        image = np.asarray(image)
        img = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
        img_resize = cv2.resize(img, dsize=(460, 460),interpolation=cv2.INTER_CUBIC)
        #img_resize = cv2.resize(image, dsize=(460, 460),interpolation=cv2.INTER_CUBIC)
        #st.text('--img_resize--')
        #st.text(img_resize)

        img_reshape = img_resize[np.newaxis,...]
        img_reshapex = resnet.preprocess_input(img_reshape)
        #st.text('--img_reshapex--')
        #st.text(img_reshapex)
```

Figura 57. Fragmento de Código Principal – Parte 2

```
        #prediction = model.predict(img_reshapex)
        #pred_class=decode_predictions(prediction,top=1)
        classes_dir = ["Adenocarcinoma","Large cell carcinoma","Normal","Squamous cell carcinoma"]
        pred_class = np.argmax(model.predict(img_reshapex))

        return classes_dir[pred_class]

if file is None:
    st.text("Por favor sube un archivo de imagen")
else:
    image = Image.open(file)
    predictions = upload_predict(image, model)
    image_class = str(predictions)
    st.write("La imagen se clasifica como",image_class)
    st.success(image_class)
    st.image(image, use_column_width=True)
```

Figura 58. Fragmento de Código Principal – Parte 3

Pruebas unitarias y redirección continua

En esta etapa, realizamos pruebas exhaustivas utilizando el conjunto de datos para evaluar el rendimiento y la precisión del modelo. Analizamos cómo el modelo se desempeña en situaciones reales y evaluamos su capacidad para hacer predicciones precisas. También consideramos otros aspectos como el tiempo de ejecución y la eficiencia computacional. Estas pruebas nos ayudan a identificar problemas y realizar ajustes antes de implementar el modelo en un entorno de producción.

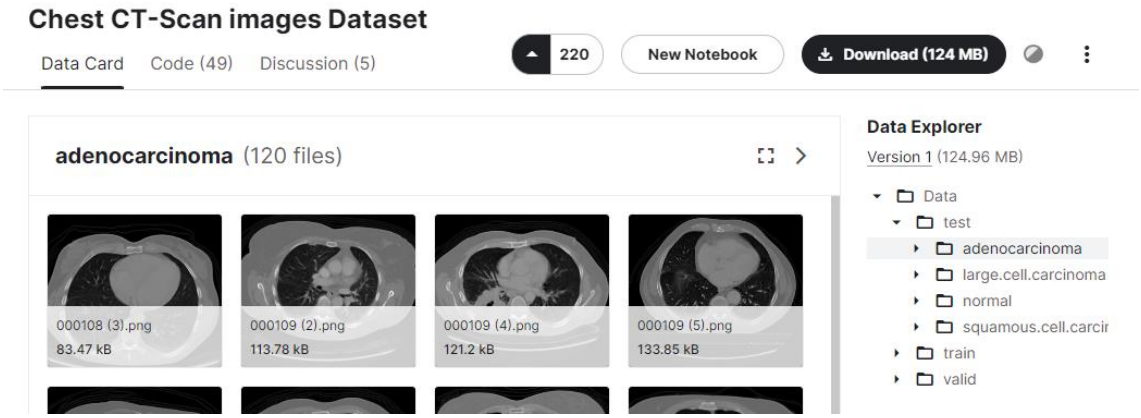


Figura 59. Imágenes del conjunto de datos

3.3.6.4. Pruebas

Pruebas de Adaptación: Para las pruebas finales, llevamos a cabo la ejecución del programa dentro de un entorno local con el objetivo de evaluar exhaustivamente su funcionalidad y rendimiento antes de proceder a la publicación de la aplicación utilizando Streamlit.

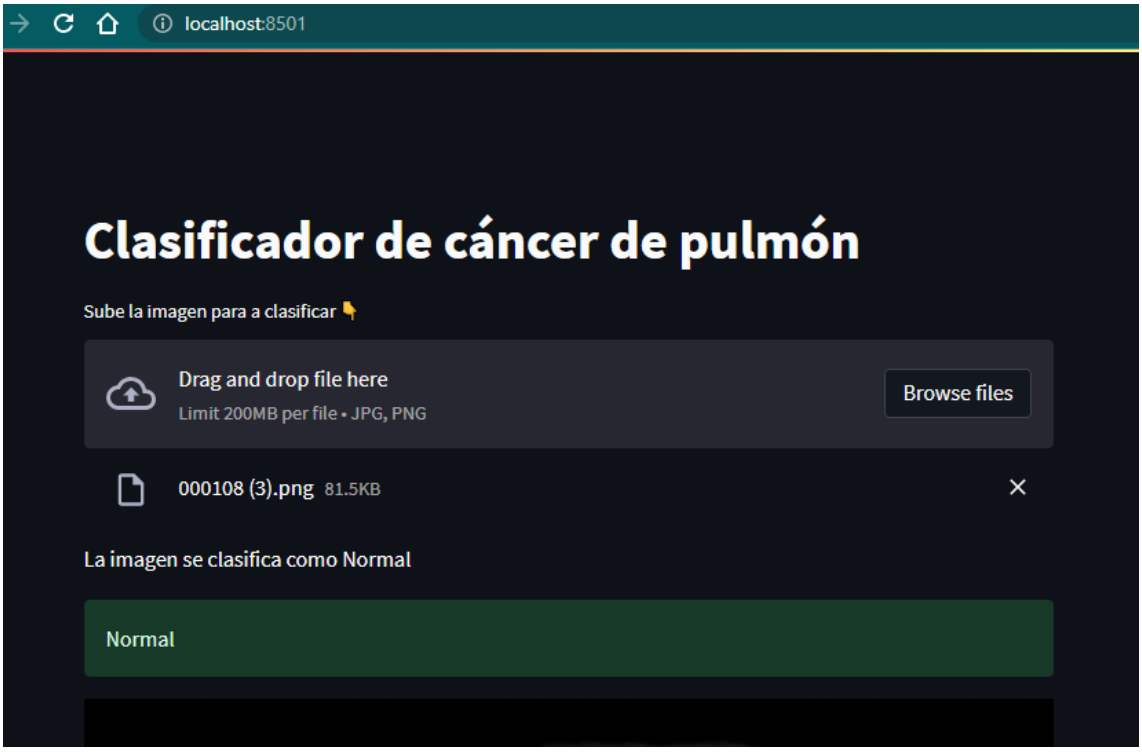


Figura 60. Pruebas en el entorno de desarrollo

3.3.6.5. Lanzamiento

Incremento del software: En la etapa de producción, utilizamos Streamlit

para subir nuestro archivo y desplegar nuestra aplicación. Esto implica proporcionar el nombre del repositorio, la rama en la que estamos trabajando y el archivo principal. Una vez realizado esto, esperamos a que se realice la configuración previa y nuestra aplicación estará disponible para el público en el siguiente enlace: https://carloswrc6-applungcancer-applungcancerhide-scdxyu.streamlit.app/

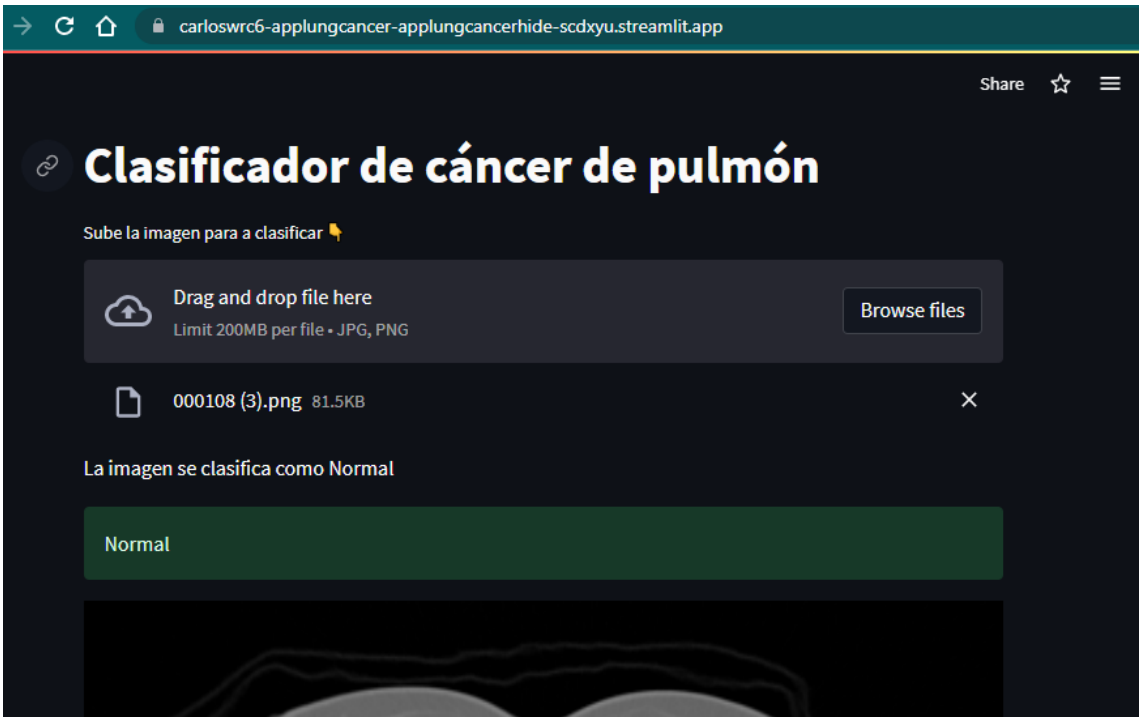


Figura 61. Entorno de Producción

IV. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

Después de comparar entre distintos conjuntos de datos el más adecuado es de tomografía computarizada de tórax “Chest CT-Scan” por qué está organizado para entrenar, validar y testear así mismo para clasificar por tipo de cáncer y es de acceso libre.

Se probaron las arquitecturas más conocidas y se optó por seleccionar Mobilenetv2, Densenet201, Efficientb4 y Resnet50 las cuales obtienen mejores resultados en la fase de entrenamiento descartando al resto de modelos.

El entorno de trabajo y lenguaje de programación seleccionados son Google Colab y Python respectivamente por la factibilidad y eficiencia con la que se puede desarrollar todo el proceso de implementación, pruebas y despliegue del sistema.

Después de evaluar las arquitecturas en relación al consumo de recursos el promedio de tiempo de respuesta de resnet50_DA es de 35min, efficientb4_DA 1h 4min 16s, densenet201_DA 34min 3s, mobilenetv2_DA 57min 7s, resnet50 18min 8s, efficientb4 26min 9s, densenet201 21min 31s y mobilenetv2 18min, en relación al consumo de GPU todos los modelos consumen 8.86GB excepto mobilenetv2 con 8.89GB, frente al consumo de memoria en GB se obtiene 2.75, 2.87, 3.17, 3.61, 2.82, 3.3, 4.42, 2.66 respectivamente, evaluando las métricas de los modelos en relación a la exactitud se obtiene 95.08%, 95.32%, 95.00%, 91.43%, 91.43%, 93.02%, 88.10% y 54.92% respectivamente, en precisión se obtiene 91.48%, 91.29%, 92.00%, 85.08%, 83.28%, 87.21%, 77.97% y 0.07% respectivamente, en sensibilidad se obtiene 88.57%, 89.84%, 87.62%, 79.68%, 82.22%, 84.44%, 73.02% y 0.06% respectivamente, en puntuación f se obtiene 89.99%, 90.54%, 89.79%, 82.38%, 82.71%, 85.77%, 75.34%, 0.07% respectivamente para cada arquitectura.

Se desarrollo una app web utilizando el streamlit lo que permitió la interacción de los usuarios comunes con el sistema.

4.2. Recomendaciones

Se recomienda la elaboración de un propio conjunto de datos disponible de imágenes integrado por instituciones públicas o privadas que considere los tipos de cáncer, estadio del mismo, número de pacientes, cantidad de imágenes y tipo de imágenes médicas el cual no deberá ser tan amplio para evitar el sobre ajuste de los modelos.

Para realizar el entrenamiento, validación y pruebas sobre nuestros modelos se deberá considerar la combinación de diversos conjuntos de datos y evaluar el rendimiento de las arquitecturas sobre las métricas establecidas.

Las investigaciones consideran clasificar la malignidad o el tipo especifico de cáncer no obstante es necesario considerar el estadio en el que se encuentra el cáncer de pulmón (I, II, III, IV) para tratar adecuadamente al paciente.

En el desarrollo de trabajos futuros se recomienda considerar que los sistemas de predicción del cáncer incluyan el cálculo del tiempo que tarde en clasificar cierta cantidad de imágenes ingresadas.

REFERENCIAS

[1] L. Revilla, “SITUACION DEL CÁNCER EN EL PERÚ, 2021,” Lima, 2021. [Online]. Available: https://www.dge.gob.pe/portal/docs/tools/teleconferencia/2021/SE252021/03.pdf

[2] Organización Panamericana de la Salud, “OPS. Epidemiología del Cáncer de pulmón en las Américas, 2014 - OPS/OMS | Organización Panamericana de la Salud,” 2014. [Online]. Available: https://www.paho.org/es/documentos/ops-epidemiologia-cancer-pulmon-americas-2014

[3] American Cancer Society, “Cáncer de pulmón,” 2021. https://www.cancer.org/es/cancer/cancer-de-pulmon.html

[4] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, “Lung Cancer Detection using CT Scan Images,” in *Procedia Computer Science*, Elsevier B.V., Jan. 2018, pp. 107–114. doi: 10.1016/j.procs.2017.12.016.

[5] X. Huang, Q. Lei, T. Xie, Y. Zhang, Z. Hu, and Q. Zhou, “Deep Transfer Convolutional Neural Network and Extreme Learning Machine for lung nodule diagnosis on CT images,” *Knowl Based Syst*, vol. 204, p. 106230, Sep. 2020, doi: 10.1016/J.KNOSYS.2020.106230.

[6] W. Abdul, “An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network,” *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, vol. 2020-December, pp. 443–446, Dec. 2020, doi: 10.1109/DESE51703.2020.9450778.

[7] M. Firdaus Abdullah, S. Noraini Sulaiman, M. Khusairi Osman, N. K. A. Karim, I. Lutfi Shuaib, and M. Danial Irfan Alhamdu, “Classification of Lung Cancer Stages from CT Scan Images Using Image Processing and k-Nearest Neighbours,” *2020 11th IEEE Control and System Graduate Research Colloquium, ICSGRC 2020 - Proceedings*, pp. 68–72, Aug. 2020, doi: 10.1109/ICSGRC49013.2020.9232492.

[8] A. Pradhan, B. Sarma, and B. K. Dey, “Lung Cancer Detection using 3D Convolutional Neural Networks,” *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, pp. 765–770, Jul. 2020, doi: 10.1109/COMPE49325.2020.9200176.

[9] A. Krishna, P. C. S. Rao, and C. Zeelan Basha, “Efficient computerized lung cancer detection using bag of words,” *2020 7th International Conference on Smart Structures and Systems, ICSSS 2020*, Jul. 2020, doi: 10.1109/ICSSS49621.2020.9202039.

[10] S. Sasikumar, P. N. Renjith, K. Ramesh, and K. S. Sankaran, “Attention based recurrent neural network for lung cancer detection,” *Proceedings of the 4th*

International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020, pp. 720–724, Oct. 2020, doi: 10.1109/I-SMAC49090.2020.9243556.

[11] C. Wang *et al.*, “Deep learning for predicting subtype classification and survival of lung adenocarcinoma on computed tomography,” *Transl Oncol*, vol. 14, no. 8, p. 101141, Aug. 2021, doi: 10.1016/J.TRANON.2021.101141.

[12] Q. Firdaus, R. Sigit, T. Harsono, and A. Anwar, “Lung cancer detection based on ct-scan images with detection features using gray level co-occurrence matrix (glcm) and support vector machine (svm) methods,” *IES 2020 - International Electronics Symposium: The Role of Autonomous and Intelligent Systems for Human Life and Comfort*, pp. 643–648, Sep. 2020, doi: 10.1109/IES50839.2020.9231663.

[13] A. Hoque, A. K. M. A. Farabi, F. Ahmed, and M. Z. Islam, “Automated Detection of Lung Cancer Using CT Scan Images,” *2020 IEEE Region 10 Symposium, TENSYMP 2020*, pp. 1030–1033, Jun. 2020, doi: 10.1109/TENSYMP50017.2020.9230861.

[14] A. Rehman, M. Kashif, I. Abunadi, and N. Ayesha, “Lung Cancer Detection and Classification from Chest CT Scans Using Machine Learning Techniques,” *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, pp. 101–104, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425269.

[15] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, “Lung Cancer Detection Using Image Processing and Machine Learning HealthCare,” *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, Nov. 2018, doi: 10.1109/ICCTCT.2018.8551008.

[16] A. M. Rossetto and W. Zhou, “Deep Learning for Categorization of Lung Cancer CT Images,” *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*, pp. 272–273, Aug. 2017, doi: 10.1109/CHASE.2017.98.

[17] S. M. Salaken, A. Khosravi, A. Khatami, S. Nahavandi, and M. A. Hosen, “Lung cancer classification using deep learned features on low population dataset,” *Canadian Conference on Electrical and Computer Engineering*, Jun. 2017, doi: 10.1109/CCECE.2017.7946700.

[18] R. Y. Bhalerao, H. P. Jani, R. K. Gaitonde, and V. Raut, “A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks,” *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, pp. 577–583, Mar. 2019, doi: 10.1109/ICACCS.2019.8728348.

[19] E. Leo. Portiansky, “Análisis multidimensional de imágenes digitales,” p. 382, 2013.

[20] J. P. Alvarado Moya, “Procesamiento y Análisis de Imágenes Digitales,” Cartago,

May 2012. [Online]. Available: https://www.tec.ac.cr/sites/default/files/media/doc/paid.pdf

[21] R. Benítez Iglesias, *Inteligencia artificial avanzada*. Barcelona: UOC, 2013. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=cat04902a&AN=mon.152595&site=eds-live

[22] V. Roman, “Introducción al Machine Learning: Una Guía Desde Cero,” Feb. 06, 2019. https://medium.com/datos-y-ciencia/introduccion-al-machine-learning-una-gu%C3%ADa-desde-cero-b696a2ead359

[23] L. Gonzalez, “K Vecinos más Cercanos - Teoría,” Apr. 05, 2019. https://aprendeia.com/algoritmo-k-vecinos-mas-cercanos-teoria-machine-learning/

[24] Jordi. Casas Roma, Anna. Bosch Rué, and Toni. Lozano Bagén, *Deep learning : principios y fundamentos*. 2019.

[25] M. Shuttleworth, “Diseño Cuasi-Experimental,” Aug. 13, 2008. https://explorable.com/es/diseno-cuasi-experimental

[26] H. Sánchez Carlessi, C. Romero Reyes, and K. Mejía Sáenz, “Manual de términos en investigación científica, tecnológica y humanística,” Lima, 2018. [Online]. Available: https://www.urp.edu.pe/pdf/id/13350/n/libro-manual-%09de-terminos-en-investigacion.pdf

[27] Google, “Introducción al aprendizaje automático | Machine Learning | Google Developers,” 2021. https://developers.google.com/machine-learning/crash-course/ml-intro?hl=es-419

[28] *Ley De Protección De Datos Personales*. Lima: CONGRESO DE LA REPÚBLICA, 2011.

[29] Colegio de ingenieros del Perú, “CÓDIGO DEONTOLÓGICO DEL COLEGIO DE INGENIEROS DEL PERÚ,” Huancayo, 2012. [Online]. Available: https://www.cip.org.pe/publicaciones/2018/CODIGO_DEONTOLOGICO2012.pdf

[30] M. I. Ullah and S. K. Kuri, “Lung nodule Detection and Classification using Deep Neural Network,” *2020 IEEE Region 10 Symposium, TENSYMP 2020*, pp. 1062–1065, Jun. 2020, doi: 10.1109/TENSYMP50017.2020.9230793.

[31] M. Islam, A. H. Mahamud, and R. Rab, “Analysis of CT Scan Images to Predict Lung Cancer Stages Using Image Processing Techniques,” *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019*, pp. 961–967, Oct. 2019, doi: 10.1109/IEMCON.2019.8936175.

[32] K. Xiang, J. Ye, and B. Xing, “Applying machine learning to facilitate personalized medicine in lung adenocarcinoma,” *PervasiveHealth: Pervasive Computing*

Technologies for Healthcare, pp. 9–12, Mar. 2019, doi: 10.1145/3314545.3314552.

[33] S. Garg and S. Garg, “Prediction of lung and colon cancer through analysis of histopathological images by utilizing Pre-trained CNN models with visualization of class activation and saliency maps,” pp. 38–45, Dec. 2020, doi: 10.1145/3442536.3442543.

[34] Z. Lin, J. Zheng, and W. Hu, “Using 3D convolutional networks with shortcut connections for improved lung nodules classification,” *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, pp. 42–49, May 2020, doi: 10.1145/3404512.3404525.

[35] J. Li, H. Chen, Y. Li, and Y. Peng, “A novel network based on densely connected fully convolutional networks for segmentation of lung tumors on multi-modal MR images,” *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, Oct. 2019, doi: 10.1145/3358331.3358400.

[36] L. Bao, T. Bao, Y. Zheng, and J. Xia, “A Simple Residual Network for Lung Nodule Classification,” *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, Jul. 2020, doi: 10.1145/3403782.3403808.

[37] Q. Wang, Y. Zhou, W. Ding, Z. Zhang, K. Muhammad, and Z. Cao, “Random Forest with Self-Paced Bootstrap Learning in Lung Cancer Prognosis,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, Apr. 2020, doi: 10.1145/3345314.

[38] S. Tripathi and S. K. Singh, “Cell Nuclei Classification in Histopathological Images using Hybrid OLConvNet,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, Mar. 2020, doi: 10.1145/3345318.

[39] L. Devnath, S. Luo, P. Summons, and D. Wang, “Performance comparison of deep learning models for black lung detection on chest X-ray radiographs,” *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, pp. 152–154, Jan. 2020, doi: 10.1145/3378936.3378968.

[40] Z. Liu, R. Wang, W. Zhang, and D. Tang, “An Unsupervised Feature Learning Method for Enhancing the Generalization of Cancer Diagnosis,” *ACM International Conference Proceeding Series*, pp. 252–257, Feb. 2021, doi: 10.1145/3457682.3457720.

[41] K. Karthick, S. Rajkumar, N. Selvanathan, U. K. B. Saravanan, M. Murali, and B. Dhiyanesh, “Analysis of Lung Cancer Detection Based on the Machine Learning Algorithm and IOT,” *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*, Jul. 2021, doi:

10.1109/ICCES51350.2021.9489084.

[42] E. D'Arnese, E. Del Sozzo, A. Chiti, T. Berger-Wolf, and M. D. Santambrogio, "Automating Lung Cancer Identification in PET/CT Imaging," *IEEE 4th International Forum on Research and Technologies for Society and Industry, RTSI 2018 - Proceedings*, Nov. 2018, doi: 10.1109/RTSI.2018.8548388.

[43] M. Jayalaxmi, J. Dhanaselvam, R. Swathi, and M. Babu, "Classification of lung nodules with feature extraction using CT scan images," *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017*, pp. 2146–2151, Jun. 2018, doi: 10.1109/ICPCSI.2017.8392097.

[44] C. H. Hsu, G. Manogaran, P. Panchatcharam, and S. Vivekanandan, "A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classifiers," *Proceedings - 8th IEEE International Symposium on Cloud and Services Computing, SC2 2018*, pp. 111–115, Dec. 2018, doi: 10.1109/SC2.2018.00023.

[45] V. G. Kale and V. B. Malode, "Lung Cancer Analysis and Diagnosis by Coalition of Photo Metric and Quality Metric Parameters," *2020 International Conference on Emerging Smart Computing and Informatics, ESCI 2020*, pp. 1–6, Mar. 2020, doi: 10.1109/ESCI48226.2020.9167532.

[46] H. Sathyan and J. V. Panicker, "Lung Nodule Classification Using Deep ConvNets on CT Images," *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, Oct. 2018, doi: 10.1109/ICCCNT.2018.8494084.

[47] A. Sivasangari, "High performance classification algorithm for analyzing medical images through computer aided diagnosis," *Mater Today Proc*, Dec. 2020, doi: 10.1016/J.MATPR.2020.11.063.

[48] Z. Liu, C. Yao, H. Yu, and T. Wu, "Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things," *Future Generation Computer Systems*, vol. 97, pp. 1–9, Aug. 2019, doi: 10.1016/J.FUTURE.2019.02.068.

[49] B. Karthiga and M. Rekha, "Feature extraction and I-NB classification of CT images for early lung cancer detection," *Mater Today Proc*, vol. 33, pp. 3334–3341, Jan. 2020, doi: 10.1016/J.MATPR.2020.04.896.

[50] S. Wang *et al.*, "ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network," *EBioMedicine*, vol. 50, pp. 103–110, Dec. 2019, doi: 10.1016/J.EBIOM.2019.10.033.

[51] X. Zhao *et al.*, "A cross-modal 3D deep learning for accurate lymph node metastasis prediction in clinical stage T1 lung adenocarcinoma," *Lung Cancer*, vol. 145, pp. 10–

17, Jul. 2020, doi: 10.1016/J.LUNGCAN.2020.04.014.

[52] H. Chen, M. Liang, X. Li, T. Wu, L. Zhang, and X. Liu, “An individualised radiomics composite model predicting prognosis of stage 1 solid lung adenocarcinoma,” *Clin Radiol*, vol. 75, no. 7, pp. 562.e11-562.e19, Jul. 2020, doi: 10.1016/J.CRAD.2020.03.019.

[53] A. Masood *et al.*, “Automated Decision Support System for Lung Cancer Detection and Classification via Enhanced RFCN with Multilayer Fusion RPN,” *IEEE Trans Industr Inform*, vol. 16, no. 12, pp. 7791–7801, Dec. 2020, doi: 10.1109/TII.2020.2972918.

[54] P. Lobo and S. Guruprasad, “Classification and Segmentation Techniques for Detection of Lung Cancer from CT Images,” *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, pp. 1014–1019, Dec. 2018, doi: 10.1109/ICIRCA.2018.8597273.

[55] R. Tekade and K. Rajeswari, “Lung Cancer Detection and Classification Using Deep Learning,” *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, Jul. 2018, doi: 10.1109/ICCUBEA.2018.8697352.

[56] O. Ozdemir, R. L. Russell, and A. A. Berlin, “A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans,” *IEEE Trans Med Imaging*, vol. 39, no. 5, pp. 1419–1429, May 2020, doi: 10.1109/TMI.2019.2947595.

[57] N. Mohanapriya, B. Kalaavathi, and T. S. Kuamr, “Lung Tumor Classification and Detection from CT Scan Images using Deep Convolutional Neural Networks (DCNN),” *Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019*, pp. 800–805, Dec. 2019, doi: 10.1109/ICCIKE47802.2019.9004247.

[58] S. Pang, Y. Zhang, M. Ding, X. Wang, and X. Xie, “A Deep Model for Lung Cancer Type Identification by Densely Connected Convolutional Networks and Adaptive Boosting,” *IEEE Access*, vol. 8, pp. 4799–4805, 2020, doi: 10.1109/ACCESS.2019.2962862.

[59] M. S. Rahman, P. C. Shill, and Z. Homayra, “A New Method for Lung Nodule Detection Using Deep Neural Networks for CT Images,” *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, Apr. 2019, doi: 10.1109/ECACE.2019.8679439.

[60] X. Huang, J. Shan, and V. Vaidya, “Lung nodule detection in CT using 3D convolutional neural networks,” *Proceedings - International Symposium on*

Biomedical Imaging, pp. 379–383, Jun. 2017, doi: 10.1109/ISBI.2017.7950542.

[61] G. Cao, T. Huang, K. Hou, W. Cao, P. Liu, and J. Zhang, “3D Convolutional Neural Networks Fusion Model for Lung Nodule Detection onClinical CT Scans,” *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 973–978, Jan. 2019, doi: 10.1109/BIBM.2018.8621468.

[62] M. B. Khumancha, A. Barai, and C. B. R. Rao, “Lung Cancer Detection from Computed Tomography (CT) Scans using Convolutional Neural Network,” *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, Jul. 2019, doi: 10.1109/ICCCNT45670.2019.8944824.

[63] I. Ali, M. Muzammil, I. U. Haq, A. A. Khaliq, and S. Abdullah, “Efficient lung nodule classification using transferable texture convolutional neural network,” *IEEE Access*, vol. 8, pp. 175859–175870, 2020, doi: 10.1109/ACCESS.2020.3026080.

[64] S. Zheng, J. Guo, X. Cui, R. N. J. Veldhuis, M. Oudkerk, and P. M. A. Van Ooijen, “Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection,” *IEEE Trans Med Imaging*, vol. 39, no. 3, pp. 797–805, Mar. 2020, doi: 10.1109/TMI.2019.2935553.

[65] A. Krishna, P. C. Srinivasa Rao, and Z. Basha, “Computerized Classification of CT Lung Images using CNN with Watershed Segmentation,” *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020*, pp. 18–21, Jul. 2020, doi: 10.1109/ICIRCA48905.2020.9183203.

[66] A. Sreekumar, K. R. Nair, S. Sudheer, H. Ganesh Nayar, and J. J. Nair, “Malignant Lung Nodule Detection using Deep Learning,” *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, pp. 209–212, Jul. 2020, doi: 10.1109/ICCSP48568.2020.9182258.

[67] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, “Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks,” *Measurement*, vol. 145, pp. 702–712, Oct. 2019, doi: 10.1016/J.MEASUREMENT.2019.05.027.

[68] D. Chen, D. Chen, J. Zhang, K. Wang, D. Qian, and X. Dong, “Segmentation of Lung Adenocarcinoma Cells Pathological Image Based on Deep Learning Method,” *ACM International Conference Proceeding Series*, pp. 226–231, Jan. 2021, doi: 10.1145/3447587.3447621.

[69] S. Mehmood *et al.*, “Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning with Class Selective Image Processing,” *IEEE Access*, vol. 10, pp. 25657–25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[70] L. Tiwari *et al.*, “Detection of lung nodule and cancer using novel Mask-3 FCM and TWEDLNN algorithms,” *Measurement*, vol. 172, p. 108882, Feb. 2021, doi: 10.1016/J.MEASUREMENT.2020.108882.

[71] S. Marques, F. Schiavo, C. A. Ferreira, J. Pedrosa, A. Cunha, and A. Campilho, “A multi-task CNN approach for lung nodule malignancy classification and characterization,” *Expert Syst Appl*, vol. 184, p. 115469, Dec. 2021, doi: 10.1016/J.ESWA.2021.115469.

[72] H. Jiang, F. Shen, F. Gao, and W. Han, “Learning efficient, explainable and discriminative representations for pulmonary nodules classification,” *Pattern Recognit*, vol. 113, p. 107825, May 2021, doi: 10.1016/J.PATCOG.2021.107825.

[73] M. M. Naeem Abid, T. Zia, M. Ghafoor, and D. Windridge, “Multi-view Convolutional Recurrent Neural Networks for Lung Cancer Nodule Identification,” *Neurocomputing*, vol. 453, pp. 299–311, Sep. 2021, doi: 10.1016/J.NEUCOM.2020.06.144.

[74] C. F. J. Kuo, J. Barman, C. W. Hsieh, and H. H. Hsu, “Fast fully automatic detection, classification and 3D reconstruction of pulmonary nodules in CT images by local image feature analysis,” *Biomed Signal Process Control*, vol. 68, p. 102790, Jul. 2021, doi: 10.1016/J.BSPC.2021.102790.

[75] M. Tsivgoulis, T. Papastergiou, and V. Megalooikonomou, “An improved SqueezeNet model for the diagnosis of lung cancer in CT scans,” *Machine Learning with Applications*, vol. 10, p. 100399, Dec. 2022, doi: 10.1016/J.MLWA.2022.100399.

[76] J. Maruthi Nagendra Prasad, S. Chakravarthy, and M. Vamsi Krishna, “A novel approach to CAD for the detection of small cell and non-small cell lung cancers,” *Mater Today Proc*, Feb. 2021, doi: 10.1016/J.MATPR.2020.12.1064.

[77] P. Dutande, U. Baid, and S. Talbar, “LNCDS: A 2D-3D cascaded CNN approach for lung nodule classification, detection and segmentation,” *Biomed Signal Process Control*, vol. 67, p. 102527, May 2021, doi: 10.1016/J.BSPC.2021.102527.

[78] L. Sun *et al.*, “Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection,” *Comput Biol Med*, vol. 133, p. 104357, Jun. 2021, doi: 10.1016/J.COMPBIOMED.2021.104357.

[79] I. D. Apostolopoulos, N. D. Papathanasiou, and G. S. Panayiotakis, “Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning,” *Biocybern Biomed Eng*, vol. 41, no. 4, pp. 1243–1257, Oct. 2021, doi: 10.1016/J.BBE.2021.08.006.

[80] S. Jain, S. Indora, and D. K. Atal, “Lung nodule segmentation using Salp Shuffled Shepherd Optimization Algorithm-based Generative Adversarial Network,” *Comput Biol Med*, vol. 137, p. 104811, Oct. 2021, doi:

10.1016/J.COMPBIOMED.2021.104811.

[81] Y. Guo *et al.*, “Histological Subtypes Classification of Lung Cancers on CT Images Using 3D Deep Learning and Radiomics,” *Acad Radiol*, vol. 28, no. 9, pp. e258–e266, Sep. 2021, doi: 10.1016/J.ACRA.2020.06.010.

[82] X. Chen *et al.*, “A CT-based deep learning model for subsolid pulmonary nodules to distinguish minimally invasive adenocarcinoma and invasive adenocarcinoma,” *Eur J Radiol*, vol. 145, p. 110041, Dec. 2021, doi: 10.1016/J.EJRAD.2021.110041.

ANEXOS

Anexo 1. Resolución de aprobación del proyecto de investigación

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°1179--2021/FIAU-USS

Pimentel, 10 de diciembre de 2021

VISTO:
El Acta de reunión N°1611-2021 del Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS remitida mediante Oficio N°0382-2021/FIAU-IS-USS de fecha 24 de noviembre de 2021, y;

CONSIDERANDO:
Que, de conformidad con la Ley Universitaria N° 30220 en su artículo 48º que a letra dice: "La investigación constituye una función esencial y obligatoria de la universidad, que la fomenta y realiza, respondiendo a través de la producción de conocimiento y desarrollo de tecnologías a las necesidades de la sociedad, con especial énfasis en la realidad nacional. Los docentes, estudiantes y graduados participan en la actividad investigadora en su propia institución o en redes de investigación nacional o internacional, creadas por las instituciones universitarias públicas o privadas.";

Que, de conformidad con el Reglamento de grados y títulos en su artículo 21° señala: "Los temas de trabajo de investigación, trabajo académico y tesis son aprobados por el Comité de Investigación y derivados a la Facultad o Escuela de Posgrado, según corresponda, para la emisión de la resolución respectiva. El periodo de vigencia de los mismos será de dos años, a partir de su aprobación. En caso un tema perdiera vigencia, el Comité de Investigación evaluará la ampliación de la misma.

Que, de conformidad con el Reglamento de grados y títulos en su artículo 24º señala: La tesis es un estudio que debe denotar rigurosidad metodológica, originalidad, relevancia social, utilidad teórica y/o práctica en el ámbito de la escuela profesional. Para el grado de doctor se requiere una tesis de máxima rigurosidad académica y de carácter original. Es individual para la obtención de un grado; es individual o en pares para obtener un título profesional. Asimismo, en su artículo 25° señala: "El tema debe responder a alguna de las líneas de investigación institucionales de la USS S.A.C.".

Que, según documentos de Vistos el Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS acuerdan aprobar los temas de las Tesis a cargo de los estudiantes que se detallan en el anexo de la presente Resolución.

Estando a lo expuesto, y en uso de las atribuciones conferidas y de conformidad con las normas y reglamentos vigentes;

SE RESUELVE:

ARTÍCULO 1°: APROBAR, el tema de la Tesis perteneciente a la línea de investigación de INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE, a cargo de los estudiantes del Programa de estudios de INGENIERÍA DE SISTEMAS según se detalla en el anexo de la presente Resolución.

ARTÍCULO 2°: ESTABLECER, que la inscripción del Tema de la Tesis se realice a partir de emitida la presente resolución y tendrá una vigencia de dos (02) años.

ARTÍCULO 3°: DEJAR SIN EFECTO, toda Resolución emitida por la Facultad que se oponga a la presente Resolución.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE

Mg. Victor Alexci Tuesta Monteza
Decano (e) / Facultad De Ingeniería, Arquitectura Y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN SAC.

MBA. María Noelia Sialer Rivera
Secretaria Académica / Facultad de Ingeniería, Arquitectura y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN SAC.

Cc: Interesado, Archivo

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°1179--2021/FIAU-USS

Pimentel, 10 de diciembre de 2021

ANEXO

N°	AUTOR(ES)	TEMA DE TESIS
1	CABRERA SANCHEZ KEVIN ALONSO MENDOZA FERRE ESPERANZA NATALY	DESARROLLO DE UNA METODOLOGÍA DE GESTIÓN DE RIESGOS PARA MEJORAR LA DISPONIBILIDAD DE SERVICIO DE TI DE UN MUNICIPIO DISTRITAL
2	ROJAS ARRUNATEGUI JOEL ENRIQUE YAFAC LAU CESAR LEONIDAS	DESARROLLO DE UN MODELO DE PROCESOS PARA LA ADQUISICIÓN DE SOFTWARE BASADO EN LA NTP-ISO/IEC 12207 PARA MEJORAR LA GESTIÓN DE LAS ADQUISICIONES DE SOFTWARE EN MICROEMPRESAS PERUANAS
3	FERNANDEZ MALUQUIS JOSE EFRAIN	ANALISIS DE ALGORITMOS BALANCEADORES DE CARGA PARA UN CLÚSTER DE SERVIDORES PARA MEJORAR LA DISPONIBILIDAD DE UN SERVIDOR
4	RAMOS SANDOVAL FABIOLA ARACELY CANTORAL MONTEJO CESAR ENRIQUE	DESARROLLO DE UN MÉTODO DE CLASIFICACIÓN AUTOMÁTICA PARA LA DETECCIÓN EFICIENTE DEL RIESGO DE ANEMIA INFANTIL A PARTIR DE HÁBITOS DE ALIMENTACIÓN Y CUIDADOS
5	BOCANEGRA GUERRERO YERSON HUAMAN HUANCAS DERBIS	ANÁLISIS COMPARATIVO DE ARQUITECTURAS DE APRENDIZAJE PROFUNDO PARA LA CLASIFICACIÓN DE ROYA AMARILLA EN HOJAS DE CAFÉ
6	SANDOVAL CHERO CESAR ARTURO	MODELO DE LA GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN ALINEADA A LA NORMA ISO/IEC 27001 ORIENTADO A LAS MICROEMPRESAS
7	DENNIS MAURICIO AVILES ODAR	APLICACIÓN DE BUENAS PRÁCTICAS PARA ENTORNOS DE DESARROLLO DE SOFTWARE BASADOS EN DEVOPS PARA MEJORAR LA INTEGRACIÓN Y DESPLIEGUE DE PROYECTOS EN UNA EMPRESA CONSULTORA DE LA CIUDAD DE CHICLAYO
8	RIVAS PLATA CASAS CARLOS GUALBERTO	DETECCIÓN DE CÁNCER DE PULMÓN EN IMÁGENES DE TOMOGRAFÍAS MEDIANTE PROCESAMIENTO DE IMÁGENES Y APRENDIZAJE AUTOMÁTICO
9	PECHE SANCHEZ CHRISTIAN WILFREDO	DISEÑO DE ARQUITECTURA DE MICROSERVICIOS PARA OPTIMIZAR PROCESOS EN LA GESTIÓN DE VENTAS ONLINE
10	SEVERINO HERNÁNDEZ YAMPIER GILBERTO	EVALUACIÓN DEL RENDIMIENTO DE UNA APLICACIÓN WEB CON ARQUITECTURA DE MICROSERVICIOS SOPORTADOS EN LA NUBE EN UN AMBIENTE DE ALTA CONCURRENCIA
11	CHANG HIDALGO HAWARD MIGUEL	COMPARACIÓN DE TÉCNICAS DE ESTIMACIÓN BASADAS EN MACHINE LEARNING PARA PREDECIR COSTOS EN LOS PLANES DE ADQUISICIONES DE LAS ENTIDADES PÚBLICAS DEL PERÚ
12	PUICON PISFIL MIRIAN ALICIA VILCHEZ CHANGANAQUI RICHARD ALEXIS	DESARROLLO DE UN MODELO DE PROCESOS BASADO EN ESTÁNDARES PARA LA EVALUACIÓN DE LA USABILIDAD WEB PARA MICROEMPRESAS PERUANAS
13	LOPEZ ABANTO GUILLERMO ANTONIO	EVALUACIÓN DE LA SEGURIDAD DE UN SISTEMA DE VOTACIÓN ELECTRÓNICA CON BLOCKCHAIN
14	CALDERON ZUÑIGA JESUS TELLO TANTARICO DILSON GUZMAN	DESARROLLO DE UN MODELO DE GOBERNANZA DE TI BASADO EN MARCOS DE GOBIERNO Y GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN PARA INSTITUCIONES PÚBLICAS PERUANAS

UNIVERSIDAD SEÑOR DE SIPÁN
FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO
CHICLAYO - PERÚ

Anexo 2: Matriz de Confusión.

	Predicción		
Actual		**Positivo**	**Negativo**
	Positivo		
	Negativo		

Anexo 3: Indicadores de Rendimiento.

Algoritmo	**Descripción**	**Exactitud**	**Precisión**	**Recall**	**F**	**Matriz de Confusión**

Anexo 4: Indicadores de Consumo.

Algoritmo	**Descripción**	**Consumo de GPU**	**Consumo de memoria**	**Tiempo de respuesta**

Anexo 5: Cronograma de desarrollo.

Project Name **ML App Lung Cancer**

		Nombre	Duración	Inicio	Fin
1		**Proyecto - App Lung Cancer**	**15días?**	**12/01/2022**	**12/21/2022**
2		**Planificación**	**2días?**	**12/01/2022**	**12/02/2022**
3		Historias de usuario, Valores, Criterios de adaptación, Plan de Iteración	2días?	12/01/2022	12/02/2022
4		**Diseño**	**3días?**	**12/06/2022**	**12/08/2022**
5		Diseños simples, Tarjetas SRC, Prototipos	3días?	12/06/2022	12/08/2022
6		**Codificación**	**5días?**	**12/09/2022**	**12/15/2022**
7		Programación, Rediseño	4días?	12/09/2022	12/14/2022
8		Pruebas Unitarias, Redirección Continua	2días?	12/14/2022	12/15/2022
9		**Pruebas**	**1día?**	**12/16/2022**	**12/16/2022**
10		Pruebas de adaptación	1día?	12/16/2022	12/16/2022
11		**Lanzamiento**	**3días?**	**12/19/2022**	**12/21/2022**
12		Incremento de software	3días?	12/19/2022	12/21/2022