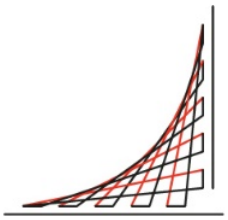
Universidad del Rosario



Escuela Colombiana de Ingeniería Julio Garavito

Detección automática de melanoma aplicando métodos de aprendizaje profundo para el procesamiento y análisis de imágenes dermatológicas

Paula Caterine Moreno Luna

Trabajo Dirigido

Tutor
Ph.D. Oscar Julián Perdomo Charry

UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ, COLOMBIA
2021

Detección automática de melanoma aplicando métodos de aprendizaje profundo para el procesamiento y análisis de imágenes dermatológicas

Paula Caterine Moreno Luna

Trabajo Dirigido presentado como requisito para aplicar al título de:
Ingeniero Biomédico

Tutor:
Ph.D. Oscar Julián Perdomo Charry

Grupo de Investigación:
Grupo de Investigación GIBIOME

UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ, COLOMBIA
2021

A mi familia por enseñarme a ser perseverante

Agradecimientos

En primer lugar, quiero darles un agradecimiento a mis papás por apoyarme y enseñarme a seguir mis sueños, por siempre creer en mi aun cuando no tienen muy claro que es lo que estudio, por inculcarme el sentido de la responsabilidad, de perseverar, de no desistir, sin ellos no sería la persona que soy hoy en día.

Quiero agradecer a mis hermanos, quienes siempre estuvieron ahí para escucharme, para ayudarme y para guiarme, en muchas ocasiones su conocimiento me enfoco en el camino para el desarrollo de diferentes tareas. A mis compañeros y amigos, que siempre apasionados por nuestra profesión dimos lo mejor de nosotros, cada actividad realizada juntos era una gran oportunidad de aprendizaje tanto como personas, como futuros ingenieros.

A su vez quiero agradecer a la Doctora Natalia Pedraza por la guía que me brindo durante la realización de este documento para la información acerca de melanoma que aquí se encuentra; a su vez, quiero agradecerle al ingeniero Oscar Julian Perdomo Charry por ser un excelente tutor durante el desarrollo de este proyecto, por estar presente ante cualquier dificultad, por la dedicación y pasión que me transmitió a mí también, muchas gracias ingeniero.

Finalmente quiero agradecerles a las dos instituciones que me acogieron a lo largo de esta carrera, quienes me brindaron el conocimiento y las experiencias necesarias para lograr culminar hoy este proyecto. Gracias por ayudarme a cumplir mi sueño.

Resumen

El melanoma, en los últimos años, ha presentado un aumento en su tasa de mortalidad, a pesar de ser el que presenta menor frecuencia de diagnóstico, convirtiéndose en un problema de salud pública tanto a nivel nacional como global, existen muchos factores de riesgo que generan una mayor probabilidad de sufrir de esta neoplasia como lo es una exposición prolongada a los rayos del sol y a los UV o tener parientes consanguíneos que lo hayan padecido.

Según El Fondo Colombiano de Enfermedades de Alto Costo, el tiempo de espera para obtener el diagnóstico de cáncer de piel tipo melanoma se encuentra entre 30 y 40 días, sin embargo, antes de este tiempo puede llegar a pasar hasta un año para que el paciente acuda a consulta médica relacionada con el aspecto anormal de un lunar; debido a esto es que existen diversos métodos de ayuda diagnóstica como la regla de ABCD que fomentan la educación para la identificación de signos de alarma.

Por esta razón el desarrollo de herramientas de apoyo diagnóstico resulta ser de gran importancia para la disminución de los tiempos y, posiblemente, de la tasa de mortalidad que está presentando este cáncer hoy en día, pues su implementación bien sea en usuarios no médicos o médicos no dermatológicos permite fomentar la educación, diferenciación e importancia de esta enfermedad en la sociedad.

El presente documento reporta el desarrollo de un trabajo dirigido que tiene como objetivo principal la aplicación de métodos basados en inteligencia artificial, específicamente, de aprendizaje profundo, para el procesamiento y análisis de imágenes dermatológicas de lunares para la detección automática de melanoma, las cuales fueron tomadas de la base de datos ISIC 2016, la cual hace parte de un reto que año a año tiene como objetivo ayudar a reducir la mortalidad de esta neoplasia, el reto del año mencionado se enfoca en diferenciar melanoma de otros trastornos de la piel y cuenta con 900 imágenes para entrenamiento, 379 para prueba, todas tomadas en distintos centros dermatológicos.

Con base en esto, se encuentra el desarrollo del marco teórico que permite dar una base de conocimiento para el entendimiento tanto del documento como del presente proyecto, seguido de esto se explica la metodología donde se desarrollan diferentes algoritmos para segmentación y clasificación, usando el lenguaje de programación Python a través de la herramienta de *Google Colaboratory*, con los cuales se desarrollan distintos modelos a ser evaluados según su desempeño con base en diferentes métricas, como índice Jaccard y Dice para segmentación y exactitud, especificidad, sensibilidad y coeficiente Kappa para clasificación.

Después se encuentran los resultados y el análisis de los resultados, donde se muestra que los modelos de segmentación presentan un muy buen desempeño, con respecto a las métricas empleadas (valores superiores a 0.60) a la hora de obtener las máscaras binarias de la predicción luego del entrenamiento, con el inconveniente de que cuando las imágenes originales presentan un contraste bajo entre la piel y el lunar, generalmente un tono azulado, la segmentación no se realiza de manera correcta.

Finalmente, para la clasificación se obtienen diferentes modelos con variaciones en sus parámetros, se observó que el desempeño que presenta cada uno de los modelos varía según la arquitectura usada permitiendo en ocasiones evitar el *overfitting*, por otro lado, el número de épocas, el optimizador y el *dropout* influyen en métricas como la especificidad para determinar la cantidad de melanomas que clasifica correctamente el modelo estudiado.

Palabras clave: Clasificación, Deep Learning, InceptionV3, Melanoma, Métricas desempeño, ResNet50, Segmentación, U-Net..

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

1 INTRODUCCIÓN

En este capítulo, se encuentra la motivación principal para realizar este proyecto, así como los objetivos asociados al mismo y una breve descripción de la organización del documento y los capítulos que lo componen.

1.1 Motivación

Según la Liga Colombiana contra el cáncer, el cáncer de piel ha aumentado su incidencia a nivel mundial, esto debido a los diferentes métodos diagnósticos que existen hoy en día y el aumento de la radiación solar por parte de toda la población mundial [8]. Se estima que un 80% de los casos diagnosticados se deben al carcinoma basocelular, seguido del escamocelular con un 15% y finalmente el melanoma con una frecuencia entre 3% - 5%, el cual, pese a que es el que presenta una menor incidencia, es responsable de aproximadamente el 80% de las muertes debidas a esta enfermedad [9], para el año 2020 se reportaron 57.043 fallecimientos en el mundo [8].

En Colombia, en el 2017 el Instituto Nacional de Cancerología reportó que el cáncer de piel es la patología tumoral maligna que presentó la mayor cantidad de diagnósticos nuevos, específicamente 712 [10]. El fondo Colombiano de Enfermedades de Alto Costo en su documento "Situación del cáncer en la población adulta atendida en el SGSSS de Colombia 2020" encontró que la prevalencia de melanoma tuvo aumento del 20% entre el 2019 y el 2020 y por otro lado, los tiempos de espera para diagnóstico e inicio del tratamiento fueron de 34 días y 77 días, respectivamente [11]. En términos de frecuencia, un estudio realizado en 2018, donde se tomaron los datos entre el periodo comprendido entre 1996 y 2010 menciona que el diagnóstico de melanoma tuvo una frecuencia del 16,1% y la incidencia para esta misma neoplasia fue de 4,6 y 4,4 por cada 100.000 habitantes para mujeres y hombres, respectivamente [9] [8].

La organización *Melanoma Research Alliance* menciona que en Estados Unidos una persona fallece cada hora todos los días a causa de melanoma, y que el riesgo de ser diagnosticado con esta neoplasia aumenta hasta un 10% al tener un familiar cercano diagnosticado [12]. Se ha demostrado que el diagnóstico temprano resulta ser de gran utilidad para mejorar el curso que toma la enfermedad, sin embargo, hay poco refuerzo de la enseñanza a los médicos sobre la identificación de este cáncer de piel, especialmente en el ambiente clínico, generando

una disminución de la facilidad que presentan los médicos y trabajadores de la salud para identificar señales de alerta, esto debido a que pocos pacientes se presentan a clínicas no especializadas cuando comienzan a tener sospechas, esto causa que el porcentaje de acierto en estos lugares sea aproximadamente del 60% [13].

En términos de económicos, en Estados Unidos se estimó que para el 2020 el costo nacional para el tratamiento de melanoma podría alcanzar los 3.16 billones de dólares [14]. En Colombia no se reporta el costo que tiene esta enfermedad para el gobierno, sin embargo, se encontró que la mayoría de los pacientes fueron tratados por el Instituto Nacional de Cancerología [15] el cual tenía con un presupuesto total de 353.110 millones para el 2021 según la Resolución 001 "Presupuesto desagregado 2021", lo que puede dar una idea acerca de los costos que representa esta enfermedad para el Estado.

Hablando acerca del descubrimiento de un lunar anormal, se evidenció que es más frecuente que el mismo paciente detecte una anormalidad en su lunar en comparación con la detección a manos de un médico, sin embargo, en términos de género, las mujeres son más propensas a detectarlo, así como el de sus parejas y pese a esto, puede pasar aproximadamente un año hasta que acudan a una evaluación médica [16].

Por estas razones mencionadas, es de vital importancia la detección temprana de melanoma tanto en Colombia como a nivel mundial, debido a esto se tiene como hipótesis que el uso de herramientas como las desarrolladas en este proyecto pueden contribuir a realizar apoyos diagnósticos y educar a las personas acerca de las señales de alerta relacionadas a esta neoplasia.

1.2 Objetivos del proyecto

1.2.1 Objetivo general

Aplicar métodos basados en inteligencia artificial para el procesamiento y análisis de imágenes dermátologicas para la detección automática de melanomas.

1.2.2 Objetivos específicos

1. Procesar y analizar imágenes de la base de datos *ISIC Challenge Datasets* 2016 empleada para la detección de melanoma.

2. Usar modelos de aprendizaje profundo para la clasificación de lunares benignos y malignos.

3. Evaluar el desempeño del modelo por medio de la obtención de diferentes métricas según cada modelo.

1.3 Organización del documento

En este documento se presentan 7 capítulos distribuidos y descritos a continuación:

- Capítulo 2: Corresponde al marco teórico donde se encuentra una selección de información de utilidad para el desarrollo de este proyecto, inicia con información relacionada al cáncer de piel haciendo un énfasis en el melanoma, sus síntomas, diagnóstico y tratamiento, y las diferentes reglas que existen para detectarlo. Posterior a esto, se hace una explicación sobre la inteligencia artificial y los diferentes métodos usados como la arquitectura U-net, ResNet50 e InceptionV3, así como de las métricas que permiten evaluar el desempeño de los modelos.

- Capítulo 3: En este capítulo se explica la metodología empleada, iniciando por la descripción de la base de datos usada, posterior a esto se explica la fase inicial, es decir, el cambio de nombre de las imágenes y la modificación del tamaño de estas. Se describe la fase de segmentación, explicando su desarrollo, entrenamiento, validación adicionalmente se explica la fase de clasificación con su respectivo desarrollo, entrenamiento y validación.

- Capítulo 4: Este capítulo corresponde a los resultados obtenidos, evidenciando las diferentes métricas de desempeño empleadas para cada una de las fases del proyecto como la sensitividad, *accuracy*, coeficiente Kappa, entre otros para la fase de clasificación, y la comparación entre las imágenes *ground truth* con las obtenidas por el modelo de segmentación acompañado del índice de Jaccard y coeficiente de Dice.

- Capítulo 5: Se presenta en análisis de resultados, donde se establecen las diferencias encontradas entre los resultados de los diferentes modelos empleados y los verdaderos positivos de la base de datos (Etiquetas y segmentación de las imágenes) estableciendo los motivos de ello mediante una revisión literaria y lo observado durante el desarrollo del proyecto. De manera adicional, se realiza una comparación con las diferentes arquitecturas usadas para explicar cómo estas son vitales para el desempeño de los modelos obtenidos.

- Capítulos 6 y 7: En los dos últimos capítulos se encuentran las conclusiones basadas en los resultados y análisis de resultados obtenidos, así como los trabajos futuros siendo este la detección en tiempo real de posible melanoma en las personas.

2 MARCO TEÓRICO

En este capítulo se exponen diferentes conceptos de utilidad para el desarrollo de este proyecto, se brinda un panorama general acerca del cáncer piel y los diferentes tipos que existen, haciendo un énfasis en el melanoma, su diagnóstico, tratamiento, factores de riesgo entre otros, adicionalmente, se hace una explicación sobre inteligencia artificial, aprendizaje profundo, también se exponen las arquitecturas empleadas así como las distintas métricas empleadas para determinar la calidad de los modelos obtenidos.

La información que se muestra a continuación surge de una revisión literaria de artículos académicos de bases de datos como Science Direct, Scopus, Google Scholar, Pubmed y SciELO.

2.1 Cáncer de piel

El cáncer de piel se entiende como un crecimiento anormal de las células que se encuentran en este tejido, una de las causas más importantes de esta enfermedad es el tiempo de latencia, o periodo de latencia, que se encuentra desde 5 hasta 40 o 45 años [17]. Por otro lado, sin importar el tipo de cáncer de piel, el principal factor de riesgo es la exposición prolongada a los rayos del sol y a los rayos UV, por esta razón la asociación de prevención dermatológica presenta cuatro recomendaciones para prevenir la aparición de esta neoplasia [18]:

- Usar protector solar dos o tres veces al día todos los días.
- Evitar la exposición prolongada a los rayos UV, especialmente entre las 11 am y las 4 pm.
- Procurar permanecer en la sombra.
- Usar ropa que contenga protección UV.

Dentro de los tres tipos de cáncer de piel, el más frecuente es el carcinoma basocelular, ubicado en la capa basal de la epidermis y los tejidos vecinos, su causa más común son las mutaciones celulares debido a la radiación ultravioleta, por esta razón se suele ubicar en áreas expuestas al sol como cara, manos o brazos, hay evidencias de que aproximadamente el 80% de cáncer basocelular se encuentra en las regiones de la cabeza y el cuello [19], sin embargo, presenta un crecimiento lento por lo que en pocas ocasiones llega a causar metástasis

en quien lo padece [20].

En Colombia, el riesgo de padecer este cáncer aumenta si se vive en el área rural, no se usan elementos de protección ante la radiación, se encuentra al aire libre gran parte del día, existen factores genéticos, características físicas de la persona (fototipos I y II, pelo y ojos claros). El diagnostico se realiza con una biopsia que parte de la sospecha del médico tratante, pues a simple vista no es sencillo de identificar; el tratamiento consiste en distintos tipos de cirugía que extraen o disecan el tumor [19].

En segundo lugar, en términos de frecuencia está el carcinoma escamocelular su origen se encuentra en la proliferación anormal de los queratinocitos suprabasales de la primera capa de la piel, su principal raíz es la exposición a la radiación UV o algunos químicos, tener fototipos I y II, padecer de albinismo, encontrarse inmunosuprimido o haber consumido tabaco [21].

La mayoría de las muertes consecuencia de los carcinomas no melanomas son a causa de el carcinoma escamocelular, debido a que este es una patología maligna que puede llegar más rápidamente a metástasis en comparación con el carcinoma basocelular, para evitar llegar a este punto es importante el diagnóstico temprano, el cual consiste en la revisión de la historia clínica así como un examen físico acompañado de ciertas preguntas relacionadas con los factores de riesgo y finalmente la confirmación por medio de una biopsia, preferiblemente amplia y con distintas capas de la piel para recopilar la información suficiente y necesaria [19].

Finalmente, el melanoma es el cáncer de piel que presenta mayor mortalidad, el riesgo de padecerlo aumenta con la edad, un diagnóstico temprano de esta neoplasia puede llegar a salvar la vida de la persona, puesto que en estados tardíos aumenta la probabilidad de fallecer por metástasis [19, 22]. Debido a que el desarrollo de este proyecto se centra en la detección automática de melanoma, esta enfermedad se explicará con mayor detalle en la siguiente sección.

2.1.1 Melanoma

El melanoma es un tumor maligno que se ubica en los melanocitos, al igual que los otros dos cánceres de piel tiene factor de riesgo asociados a factores genéticos, mutaciones genéticas, especialmente la del gen CDKN2A ubicado en el cromosoma 9, factores ambientes como la alta exposición a la radiación ultravioleta o el uso de cámaras bronceadoras [23].

Debido a las células que se ven afectadas el diagnóstico de esta neoplasia puede iniciar con el mismo paciente, para ello existen diversos métodos de diagnóstico relacionados con el aspecto del lunar:

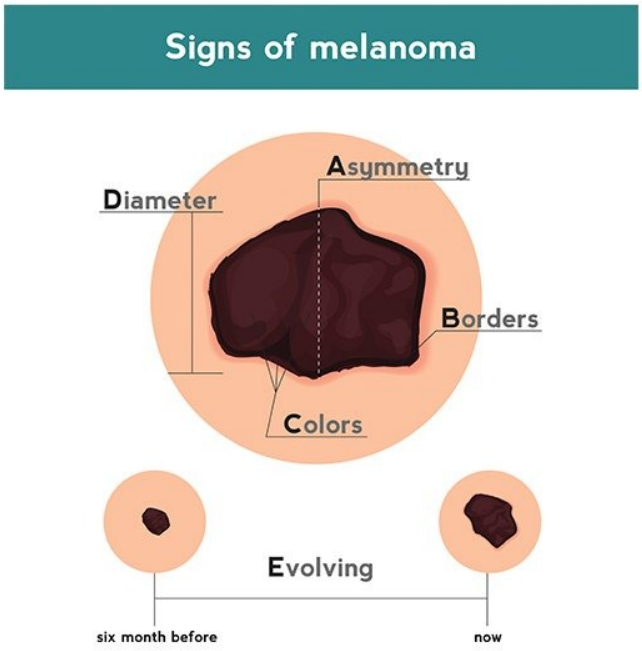


Figura 2-1: Regla ABCD para diagnóstico de melanoma. Tomada de [1]

- Método de la regla ABCD, en la imagen **2-1** se puede apreciar gráficamente lo explicado a continuación:

  1. Asimetría: Si al dividir el lunar en 2 ejes, y evaluarlo según su color, forma y estructura se evidencia que no existe asimetría se da un puntaje de 0, pero si existe asimetría en un eje o en ambos el puntaje es de 1 o 2, respectivamente [24] [25].
  2. Borde: Se divide la lesión en 8 partes aproximadamente iguales, y se obtiene un puntaje de 1 cada vez que existe una terminación brusca del borde para un puntaje máximo de 8 [24] [25].
  3. Color: Si hay presencia de más de un color en el lunar, estos pueden ser blanco, café claro u oscuro, azul, rojo o negro, se obtiene un punto por cada color para un total de 6 puntos en caso de tener presencia de los seis colores mencionados [24] [25].
  4. Diámetro: Si el diámetro es mayor a 6 mm obtiene un puntaje de 1, sin embargo, hay evidencia de melanomas que tienen un diámetro menos por lo cual actualmente se usa más para esta parte la evaluación de estructuras dermatoscópicas, conformadas por retículo pigmentado, puntos, glóbulo, entre otras [24] [25] [26].

- Lista de 7 puntos [25], en la imagen **2-2** se puede apreciar gráficamente lo explicado a continuación:

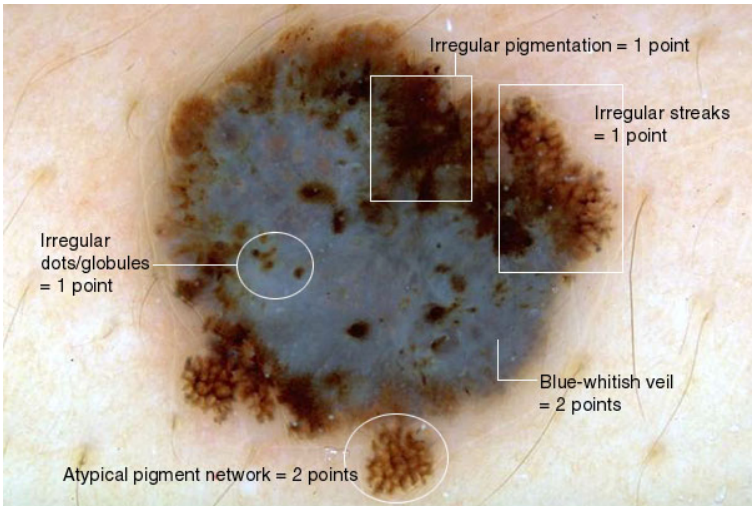


Figura 2-2: Lista de 7 puntos para diagnóstico de melanoma. Tomada de [2]

1. Retículo pigmentado: Si presenta color café, gris o negro y orificios de malla irregular y/o gruesas.
2. Vello azul-blanquecino: Zona con un tono azul que no ocupa toda la región estudiada.
3. Patrón vascular: Si hay presencia de vasos irregulares en el tumor.
4. Proyecciones irregulares: Si se encuentras proyecciones irregulares en la periferia de la lesión.
5. Puntos y/o glóbulos: Si hay de diferentes tamaños y presentan una distribución irregular.
6. Manchas de pigmento: Zonas de color café, gris o negro que tienen una distribución asimétrica a lo largo de la región estudiada.
7. Estructuras asociadas a la regresión: Si tiene áreas despigmentadas a causa de una cicatriz y/o áreas de tonos azules o grises.

Las especificaciones para calificar cada uno de los siete puntos se presentan en la tabla **2-1** [2], luego de calificar cada punto, se suman y se calcula el puntaje total, si este es igual o mayor a tres hay sospecha de presencia de melanoma, pero si, por el contrario, el puntaje es menor a tres no hay sospecha de melanoma [25].

Una vez diagnosticado, es importante reconocer las diferentes etapas que tiene este cáncer para saber la cantidad y el lugar donde se encuentra, el libro *Cutaneous Melanoma* en su capítulo 4 menciona que el melanoma cuenta con 5 etapas [27]:

- Etapa 0: Enfermedad *in situ*

Criterio	Clasificación	Puntaje máximo
Retículo pigmentado	Ausencia(0), Típico(0), Atípico(2)	2
Vello azul-blanquecino	Ausencia(0), Presencia(2)	2
Patrón vascular	Ausencia(0), Regular(0), Irregular(2)	2
Proyecciones irregulares	Ausencia(0), Regular(0), Irregular(1)	1
Puntos y/o glóbulos	Ausencia(0), Regular(0), Irregular(1)	1
Manchas de pigmento	Ausencia(0), Regular(0), Irregular(1)	1
Estructuras asociadas a la regresión	Ausencia(0), Presencia(1)	1

Tabla 2-1: Criterios para la lista de 7 puntos. Tomado de [2]

- Etapa I y II: Enfermedad localizada donde la ubicación de la ulceración permite hacer una diferenciación más detallada en esta etapa según el tamaño que presente y si tiene o no ulceración.
- Etapa III: Enfermedad regional en la cual puede haber presencia de metástasis en los ganglios linfáticos.
- Etapa IV: Melanoma en cualquier etapa con presencia de metástasis distancia generalmente en la piel, ganglios linfáticos o sitios viscerales.

El tratamiento de melanoma abarca distintas posibilidades, como el tratamiento quirúrgico para extirpar el tumor completamente con unos milímetros de seguridad para minimizar la posibilidad de volver a tener presencia de esta neoplasia, la terapia adyuvante que incluye el uso de interferón en dosis altas o bajas, la inmunoterapia la cual muestra buenos resultados en pacientes en etapa 3 de la enfermedad [28].

Finalmente, está la terapia médica, la cual se realiza cuando la posibilidad de extirpación completa presenta grandes riesgos para el paciente dentro de estas se encuentra la radioterapia que ha registrado una curación de gasta el 95% usando rayos Grenz o rayos X blandos [28].

2.2 Inteligencia artificial

La inteligencia artificial (AI por sus siglas en inglés) es la habilidad de los sistemas para adquirir y emplear el conocimiento, nacen de la inspiración del aprendizaje que realizan los humanos para ejecutar diversas tareas como detección, toma de decisiones, manipulación de objetos, entre otras. La AI ayuda al desarrollo de diferentes proyectos en muchas áreas del conocimiento, como la agricultura, el petróleo, industria textil y, como en este caso, la salud, permitiendo la automatización de robots o máquinas para la ejecución autónoma de

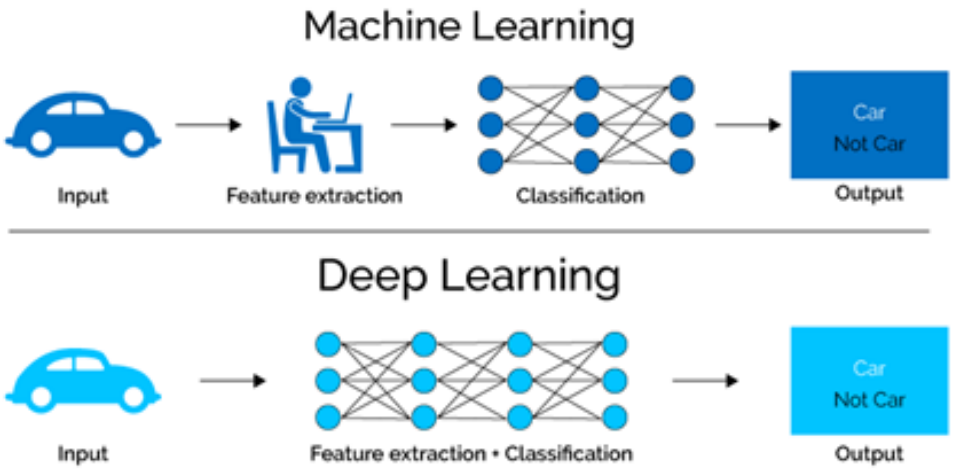


Figura 2-3: Comparación entre la extracción de características de aprendizaje de máquina y aprendizaje profundo. Tomada de [3]

sus trabajos [29].

Esta novedosa herramienta ha ayudado en el sector de la salud para el desarrollo de terapias contra el cáncer, aplicaciones para la prestación del servicio de salud, para la detección de nano-materiales en diversos tejidos del cuerpo, la identificación de características de tejidos y posibles anormalidades [30]. Y depende del objetivo del proyecto en el cual se implementa se pueden emplear los distintos tipos que existen, aprendizaje automático o aprendizaje profundo, donde este último se explica a continuación.

2.2.1 Deep Learning

Este tipo de aprendizaje no requiere de la intervención humana para la extracción de características, pues utiliza jerarquización de conceptos para construir unos más complejos, esto se ve más claro en la figura **2-3** donde son las mismas redes neuronales las que realizan la extracción de características [31]. El aprendizaje profundo es de gran utilidad para obtener características de señales como el habla, donde los rasgos para identificarlo requieren de un nivel de procesamiento muy similar al cerebro, pues lo que se desea con este tipo de inteligencia artificial es segmentar toda la información en secciones más pequeñas, permitiendo a los computadores construir conceptos complejos a partir de estas [31].

De manera particular, el uso de esta técnica de aprendizaje ha sido ampliamente usada en imágenes médicas de tipo dermatológico, por ejemplo, Abbas et al. [32] usaron Deep Learning para extraer características y clasificar imágenes demoscópicas de melanoma acral, un

tipo de melanoma que suele aparecer en las extremidades, para ello hicieron uso de diferentes modelos, el primero de ellos con ConvNet, el cual pese a ser entrenado desde cero, es decir sin un pre-entrenamiento, obtuvo resultados excelentes a la hora de clasificar las imágenes ,obtenidas del Hospital Severance en Corea del Sur, con una exactitud de 0,9103. Por otro lado, usaron ResNet y AlexNet para la misma tarea de clasificación, las cuales al evaluar su desempeño se observó que el primer modelo mencionado tuvo un mejor valor para la curva ROC, 0,1% mejor que AlexNet y 0,6% que el modelo propuesto de ConvNet, pese a esta diferencia, todos los diferentes modelos empleados en este estudio tuvieron una exactitud mayor a 0,90 lo que indica un buen desempeño general [32].

Por otro lado, este 2021, Cassidy et al. [33] emplearon las bases de datos ISIC de 2017 a 2020, la usada en el presente proyecto es de ISIC año 2016, para analizar su uso y comparar diferentes arquitecturas empleadas de aprendizaje profundo en los últimos 4 años, para ello usaron las 19 arquitecturas más usadas para clasificar entre melanoma y no melanoma. En primer lugar, usaron la base de datos ISIC 2020 y debido a que esta no cuenta con datos *ground truth* usaron la métrica de AUC que mide el área bajo la curva ROC, con ello se obtuvo el mejor desempeño para VGG19 con AUC de 0.80, por otro lado las arquitecturas que tuvieron peor desempeño (AUC menor a 0.60) fueron DenseNet169 y EfficientNetB4, el resultado de esta última se lo atribuyen al tamaño que tiene la arquitectura en comparación con el tamaño del conjunto de entrenamiento [33]. Con respecto a la evaluación de los modelos con base en los datos dados por el experto emplearon la base de datos del año 2017, en donde encontraron que VGG19 seguía manteniendo el mejor desempeño con una exactitud de 0.56, sin embargo, notaron que el *recall* mayor fue con InceptionV3, 0.94, lo que puede indicar un sobre-entrenamiento para clasificar melanoma [33].

Con base en esta información las redes de aprendizaje profundo, como las que se explicaran en las siguientes secciones, han demostrado ser de gran utilidad para la clasificación de melanoma pues al extraer las características que presentan más relación generan que su desempeño sea superior a los modelos tradicionales basados en características [32].

2.2.2 Arquitecturas

Para el desarrollo de este proyecto se emplean diferentes arquitecturas para segmentación y clasificación que han tenido buenos resultados en su uso para aplicaciones biomédicas.

U-Net

La arquitectura U-net está diseñada para tareas de segmentación, cuenta con dos caminos, uno comprimido y otro expandido; el primero de ellos es similar a la arquitectura de una red neuronal convolucional normal, es decir, consiste en la repetición de dos capas convolucionales seguidas de una rectificación lineal y una operación de muestreado donde se aumentan la

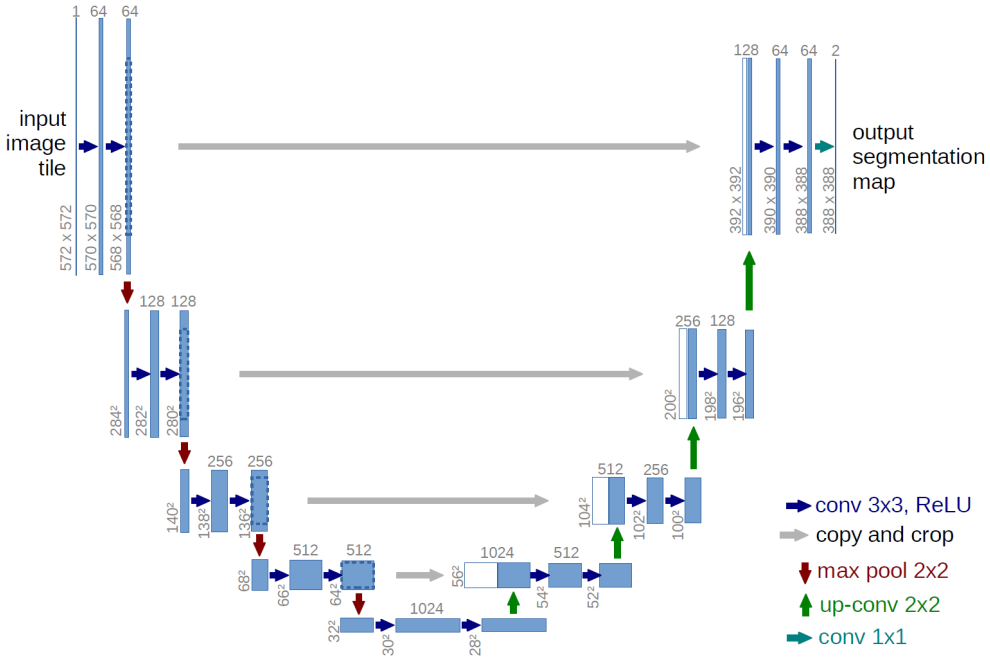


Figura 2-4: Arquitectura U-Net. Tomada de [4]

cantidad de canales que extraen características. Por otro lado, para el otro camino se inicia con la ampliación de las muestras para continuar con una convolución que disminuye los canales de características, y se finaliza con una rectificación lineal, dando como total 23 capas convolucionales; lo anterior se puede observar gráficamente en la figura **2-4** [4].

ResNet50

La arquitectura de redes neuronales residuales es una red neuronal convolucional que se diferencia en el aprendizaje de las características residuales en lugar de las características, en la figura **2-5** se puede observar que la entrada de la capa es sumada a los pesos de la salida seguido de la ejecución de una rectificación lineal, esto sucede, en este caso para las 50 capas de las que consta la arquitectura [5].

InceptionV3

Esta arquitectura es totalmente convolucional, cuenta con un módulo especial que permite mejorar el rendimiento y el desempeño de los modelos generados, internamente tiene capas convolucionales con filtros que se procesan en paralelo de forma que las características extraídas se unan antes de pasar a la siguiente capa. La gráfica de esta arquitectura se encuentra en la figura **2-6** debido a estas características y la forma como recibe y maneja la información esta arquitectura es muy útil cuando se desea evitar el *overfitting* [6].

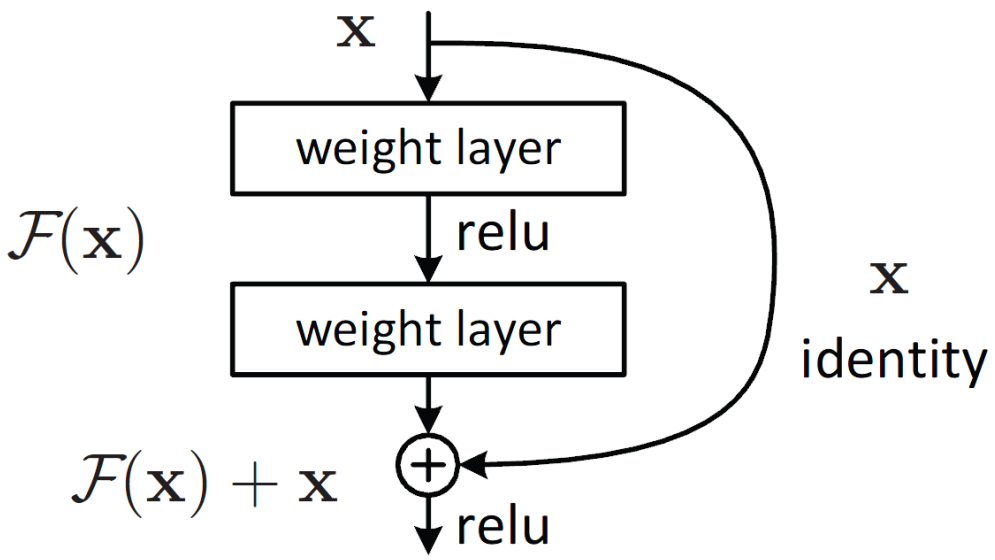


Figura 2-5: Arquitectura base de ResNet50. Tomada de [5]

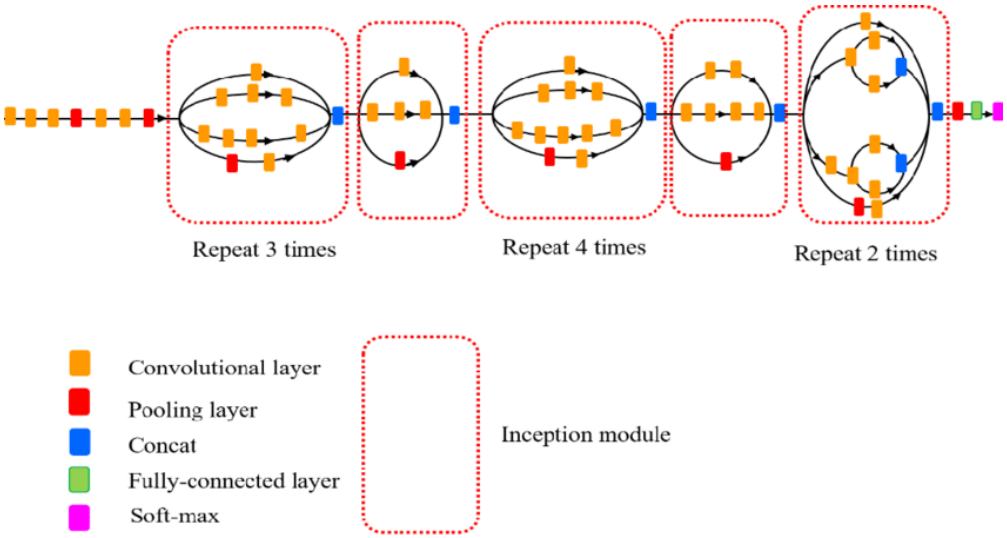


Figura 2-6: Arquitectura base de InceptionV3. Tomada de [6]

2.2.3 Métricas de desempeño

Las métricas de desempeño son importantes para observar la calidad de los modelos, así como el comportamiento que presentan estos en los conjuntos de entrenamiento vs. los conjuntos de prueba. Las expuestas a continuación fueron las empleadas para la evaluación de los modelos, estas se basan en los valores que contiene cada celda de la matriz de confusión, como se ve en la tabla **2-3**.

Estas métricas fueron seleccionadas para la evaluación debido a la importancia que tienen para determinar el desempeño y la calidad de los modelos, por ejemplo, para la segmentación los índices de Jaccard y Dice indican la similitud que presentan las máscaras binarias dadas por el modelo y las de la base de datos, entre mayor sea este índice, mayor similitud presenta, por lo que el tener valores mayores a 0,7 indican que el modelo segmentó correctamente los lunares lo que puede dar mayor seguridad a la hora de usarlas en otras aplicaciones como la clasificación con base en la regla ABCD y los parámetros dados por dermatólogos, un trabajo futuro de este proyecto.

Para la parte de clasificación se manejaron cuatro métricas diferentes, la primera de estas es la exactitud, la cual da una idea acerca de las clasificaciones totales (benignas y malignas) del modelo que fueron correctas, de manera específica se tiene la especificidad y la sensibilidad, la primera de ellas expone la cantidad de melanomas que clasificó el modelo correctamente (valor entre 0 y 1, siendo 1 una clasificación correcta), una métrica de suma importancia para el objetivo de este proyecto, la segunda métrica indica el porcentaje de lunares benignos que fueron clasificados por el modelo correctamente, todas estas métricas permiten conocer la calidad del modelo en términos de su clasificación binaria.

Finalmente, el coeficiente Kappa es una medida que indica el nivel de acuerdo que existe entre las clasificaciones dadas por los diferentes modelos y las dadas por el experto (médico dermatólogo que clasificó las imágenes en la base de datos), con base en lo observado en la tabla **2-2**, el tener un valor mayor a 0.6 indica un nivel de acuerdo bueno, es decir, una relación buena entre las clasificaciones de los modelos y las clasificaciones dadas por el experto en la base de datos.

Índice de Jaccard

El índice de Jaccard muestra la similitud entre las máscaras binarias del experto y las del modelo. Su ecuación está dada por 2-1 [34].

$$JA = \frac{TP}{TP + FN + FP} \tag{2-1}$$

Kappa	Interpretación
< 0	Pobre
0 - 0.2	Leve
0.21 - 0.4	Razonable
0.41 - 0.6	Moderado
0.61 - 0.8	Substancial
0.81 - 1	Casi perfecto

Tabla 2-2: Interpretación coeficiente Kappa. Tomado de [7]

	Verdaderas	
	TP	FP
Predecidas	FN	TN

Tabla 2-3: Matriz de confusión

Índice de Dice

El índice de Dice, similar al de Jaccard, muestra el parecido entre las imágenes del modelo, en este caso las de segmentación, y las imágenes dadas por el experto en la base de datos. Su ecuación está dada por 2-2 [34].

$$DI = \frac{2 * TP}{2 * TP + FN + FP)} \tag{2-2}$$

Exactitud

Indica el porcentaje de predicciones correctas que tuvo el modelo sobre el total de predicciones, la ecuación que lo caracteriza es 2-3 [34].

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2-3}$$

Especificidad

Indica la relación entre los verdaderos negativos y el total de negativos del modelo, su ecuación es 2-4 [34].

$$SE = \frac{TN}{TN + FP} \tag{2-4}$$

Sensibilidad

Indica la relación de los verdaderos positivos sobre el total de positivos en el modelo, su ecuación es 2-5 [34].

$$SE = \frac{TP}{TP + FN} \tag{2-5}$$

Coeficiente Kappa

Es un coeficiente de acuerdo, es decir, índica el nivel de acuerdo que existe entre el experto y el modelo, su ecuación es 2-6 [35].

$$K = \frac{2(TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN))} \tag{2-6}$$

3 METODOLOGÍA

En este capítulo se presenta la metodología implementada para la detección automática de melanoma haciendo uso de métodos de aprendizaje profundo a través de diferentes arquitecturas para segmentación y clasificación de las imágenes, seguido de la determinación de características de los lunares con base en la regla de ABCD para melanoma, incluyendo la obtención de las métricas de desempeño para analizar la calidad de los modelos, así como la comparación entre estos; lo anterior se puede apreciar de manera general en la figura **3-1**.

3.1 Base de datos

Para el desarrollo de este proyecto se usó la base de datos del reto de *"The International Skin Imaging Collaboration"* (ISIC), la cual es una asociación creada entre diferentes entidades en el mundo para colaborar y facilitar el acceso a imágenes de la piel. El objetivo de ISIC es ayudar a reducir la mortalidad causada por el melanoma al facilitar la enseñanza acerca de su reconocimiento, identificación y posible diagnóstico. La ISIC patrocina retos desde 2016 enfocados, inicialmente en la precisión de los algoritmos para diferenciar el melanoma de otros trastornos de la piel, sin embargo, a medida que avanzan los retos, año tras año estos crecen en complejidad y escala, por ejemplo, entre 2019 y 2020 se incluyó el impacto que tienen estos algoritmos en la distribución y evaluación en el ambiente clínico [36].

De manera específica se usó la base de datos del desafío de 2016, la cual cuenta con 1279 imágenes, dividas en 900 para entrenamiento (727 benignas y 173 malignas) y 379 para test (304 benignas y 75 malignas), cada una con su respectivo *ground truth* correspondiente a una máscara binaria y el *label* otorgado por un experto, siendo 0 para benigno y 1 para maligno [34].

3.2 Procesamiento y análisis de imágenes dermatológicas de lunares para la detección de melanoma

Como primer tratamiento a las imágenes se les realizó un cambio de nombre con el fin de unificar tanto las máscaras binarias como las imágenes en RGB de manera que quedaran con el mismo nombre y asociadas a la etiqueta que les corresponde, esto se puede ver de manera más clara en la tabla **3-1**, donde se aprecia el nombre asignado a la figura **3-2** del grupo de

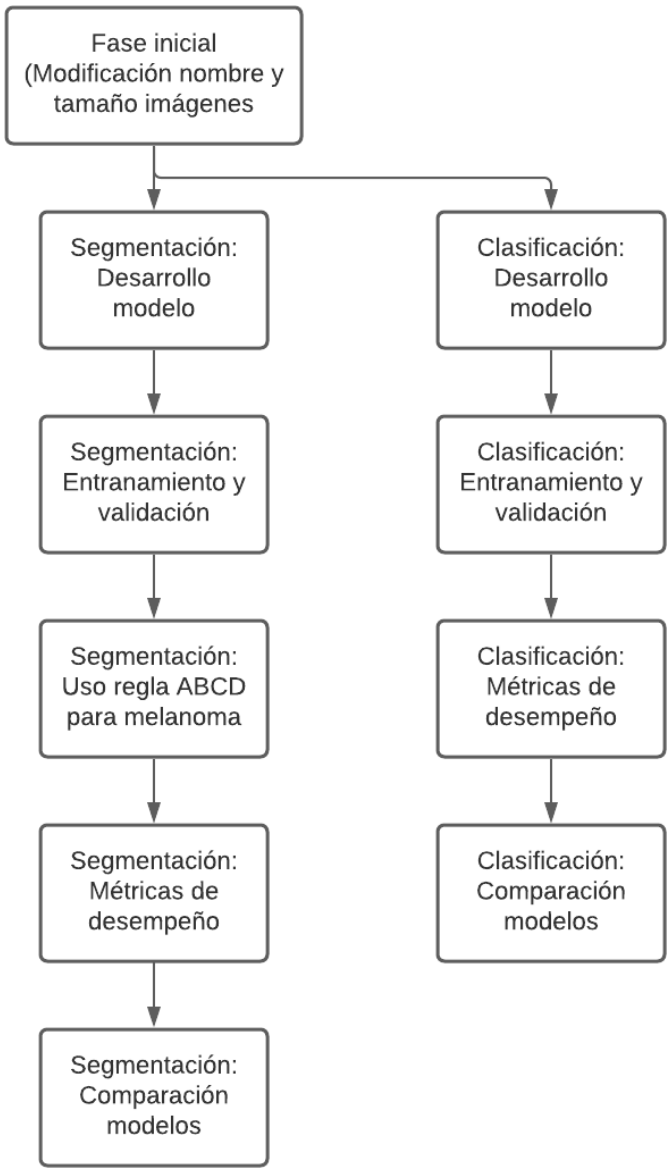


Figura 3-1: Esquema de la metodología del proyecto

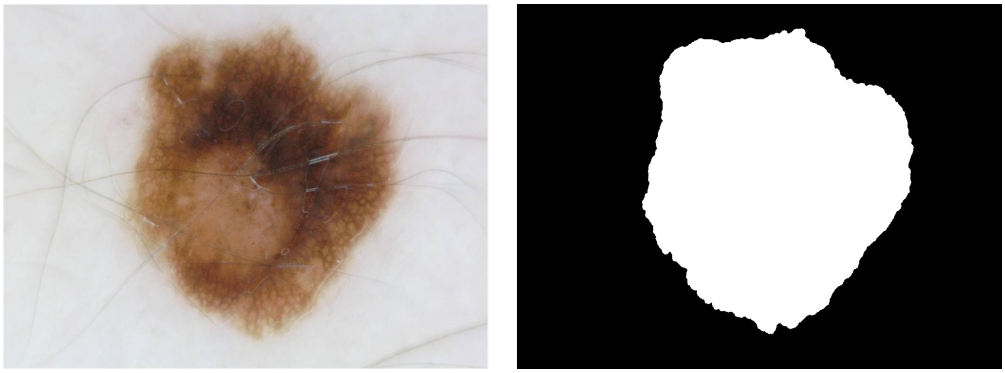


Figura 3-2: Imagen en RGB 003 de la base de datos con su respectiva máscara binaria

entrenamiento de la base de datos junto con el nombre de su máscara.

Por otro lado, se realizó una reducción en el tamaño de todas las imágenes a 224×224 pixeles, debido a que las arquitecturas empleadas admiten un tamaño específico y además para facilitar su procesamiento; lo anterior se realizó debido a varios factores que influían en el tiempo de procesamiento y desempeño de las arquitecturas, de manera detallada [37]:

- Las arquitecturas más empleadas para el aprendizaje profundo, *Deep Learning*, tienden a entrenar con valores entre 224×224, 256×256 y 299×299.

- Los modelos usados para aprendizaje profundo requieren que todas las imágenes (RGB y máscaras) tengan el mismo tamaño.

- El tamaño de las imágenes puede afectar el tiempo computacional, y el desempeño del modelo empleado.

Si bien el realizar esta acción tiene un impacto en el desempeño de las redes neuronales convolucionales, al evidenciar una disminución en métricas como exactitud [38], el tener un límite de memoria no permite el empleo de imágenes de alta resolución por lo que el realizar esta acción se consideró como la mejor opción para optimizar tiempos de computo, memoria y hacer un uso correcto de las arquitecturas empleadas.

	Original	**Renombrada**
RGB	ISIC_0000003.jpg	0_ISIC_0000003.jpg
Máscara Binaria	ISIC_0000003_Segmentation.png	0_ISIC_0000003_Segmentation.png

Tabla 3-1: Ejemplo de la modificación del nombre de las imágenes

3.3 Implementación de modelos de aprendizaje profundo para la clasificación de lunares

La implementación de los modelos de aprendizaje profundo se dividió en dos secciones explicadas en detalle a continuación:

3.3.1 Segmentación

Para la segmentación de las imágenes se hizo uso de la arquitectura U-net, la cual, como se explicó en secciones anteriores, es de gran utilidad para este tipo de tareas; para su implementación se usó principalmente el lenguaje de programación Python a través de la herramienta de Google, *Google Colaboratory*.

En primer lugar, se obtuvo la lista de las imágenes para entrenamiento y para test tanto las de RGB como las máscaras, todas estas en su nuevo tamaño, con estas listas se cargan los datos, haciendo distinción entre los dos grandes grupos ya mencionados, las cuales van a ser usadas posteriormente para el entrenamiento del modelo y la obtención de las métricas de desempeño.

Una vez ya obtenidos los datos necesarios, se inicializa el modelo a usar empezando por el uso de la red neuronal convolucional VGG-19 como *backbone*, la cual se emplea para cargar una versión pre-entrenada con las imágenes de la base de datos *ImageNet*, generando un punto de inicio para la definición del modelo. El entrenamiento se inició usando el optimizador Adam y como función de pérdida se usó *binary_crossentropy*, que calcula la perdida de entropía cruzada entre las máscaras binarias del modelo y las dadas por la base de datos. En esta parte, se realizó una exploración de hiper-parámetros tasa de aprendizaje y tamaño del *batch* guardando los pesos del modelo con menor perdida en validación usando *ModelCheckpoint*. Por otro lado, se empleó la métrica de desempeño *iou_score*, para probar el mejor modelo en entrenamiento y evaluar el conjunto de test.

Para el entrenamiento del modelo se usaron las 900 imágenes de la base de datos, con un validation split de 0.1, es decir se usaron 90 imágenes, a su vez se usó un batch size de 16.

3.3.2 Clasificación

Para el desarrollo de esta sección se usaron dos arquitecturas diferentes, ResNet50 e InceptionV3, al igual que la sección de segmentación, en *Google Colaboratory*. En esta tarea, se usaron los mismos conjuntos que en la subsección de segmentación; con estas listas obtenidas y organizadas, se cargan los datos de entrenamiento y test para su uso en las dos diferentes

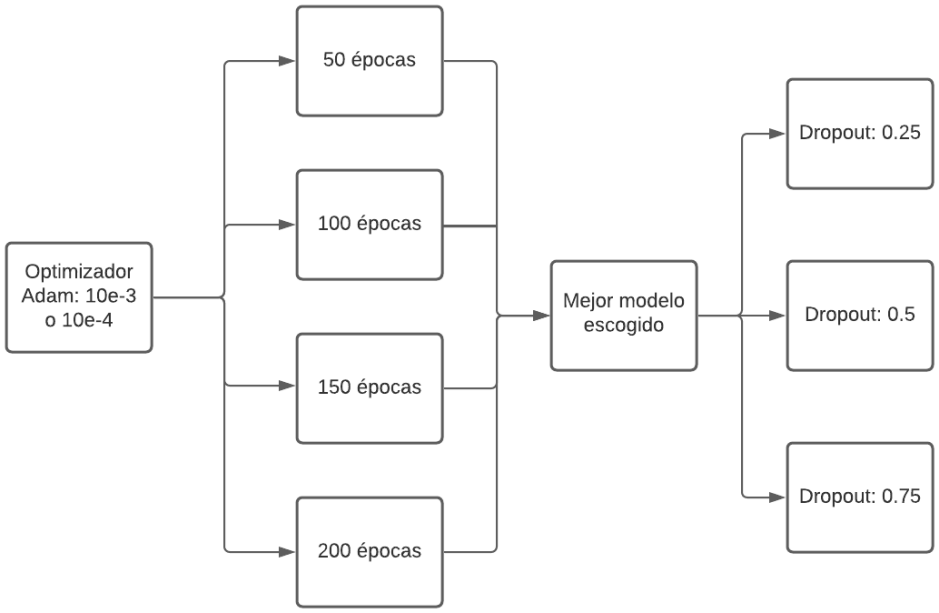


Figura 3-3: Detalle de los parámetros modificados para los distintos modelos de clasificación

arquitecturas.

Para usar ResNet50, se inicializa el modelo cargando los pesos de la base de datos *ImageNet*, también fue necesario especificar el tamaño de las imágenes y el número de clases en nuestra tarea que son 2 clases. Adicionalmente se usó la perdida en validación para guardar el mejor modelo.

Con esto ya organizado, se definen los distintos parámetros del modelo; para la primera parte se decidió no tomar en cuenta el *dropout* y variar el valor del optimizador Adam y el número de épocas por cada modelo, con esto se determinará el que mejores métricas de desempeño presente para usar este y modificar el *dropout* de forma que se pueda observar la importancia de usar distintos valores para este parámetro, lo anterior se puede apreciar de manera más clara en la figura **3-3**; el mismo procedimiento se realizó usando la arquitectura InceptionV3.

Para el entrenamiento de los modelos se usaron las 900 imágenes de la base de datos, con un validation split de 0.2, y un batch size de 16.

3.4 Evaluación del desempeño de los modelos propuestos

Como se explicó anteriormente de manera breve, una vez se obtienen los modelos entrenados y guardados se observan las métricas de desempeño para determinar la calidad de los modelos y hacer una comparación entre los distintos resultados obtenidos.

Para la parte de segmentación se obtienen las imágenes predichas por el modelo, las máscaras binarias y las originales dadas por la base de datos, a partir de esto, se obtienen las 5 mejores y las 5 peores para comparar cuantitativamente y observar cualitativamente el desempeño de nuestros modelos de segmentación. De manera cuantitativa, se obtiene el índice de Jaccard e índice de Dice de cada una de las imágenes y el promedio de cada uno de estos, y a partir de estos hacer una comparación entre el modelo con 10 épocas y el que tiene 20.

Para la parte de clasificación, para todos los distintos modelos con sus conjuntos de entrenamiento y test, se obtuvieron diferentes métricas de desempeño como lo son:

- Accuracy
- Recall o Sensitividad
- Especificidad
- Coeficiente Kappa

Para con estas hacer comparación entre todos los modelos obtenidos en esta parte del proyecto de forma que se logre identificar la importancia de cada uno de los parámetros modificados y como estos pueden afectar el desempeño y la calidad de cada uno.

4 RESULTADOS

En este capítulo se presentan los resultados obtenidos para el desarrollo del presente proyecto, donde se obtuvieron diferentes métricas de desempeño para los conjuntos de entrenamiento y test, que permitieron determinar los mejores y peores modelos, así como las mejores y peores imágenes segmentadas por el modelo junto con el índice de Jaccard y Dice. Adicionalmente, se presentan los datos obtenidos durante la fase de clasificación.

4.1 Segmentación

En primer lugar, se calculó el promedio de índice de Jaccard (IoU) de los conjuntos de entrenamiento y prueba y con base en este resultado se estimó el índice de Dice, como se puede ver en la tabla **4-1**, donde el valor máximo de IoU para el conjunto de entrenamiento fue de 0,78 (modelo con 20 épocas) y para el conjunto de test el que tuvo mejor resultado fue, de igual forma, el modelo entrenado con 20 épocas. Sin embargo, es de resaltar, que el valor de esta métrica para las 1279 imágenes de la base de datos tiende a presentar valores de IoU mayores a 0,65, lo que quiere decir que, la gran mayoría de las imágenes están teniendo una buena segmentación por parte del mejor modelo.

# épocas		Jaccard	Dice
10	**Train**	0,6466	0,7854
	Test	0,6537	0,7906
20	**Train**	0,7886	0,8818
	Test	0,7796	0,8762

Tabla 4-1: Métricas de desempeño para modelos de segmentación

Por otro lado, al observar el coeficiente de Dice, se observa que este se relaciona directamente con el índice de Jaccard, pues presentan una correlación positiva (métricas complementarias), pese a esto, a partir de Dice podemos tener un valor cercano al rendimiento del modelo, lo que para el modelo de 10 épocas significa que su rendimiento promedio se encuentra cercano al 80% mientras que el de 20 épocas, que ha presentado los mejores resultados de segmentación, este valor se encuentra más cercano al 90%. Cabe aclarar que el uso de ambos índices fue

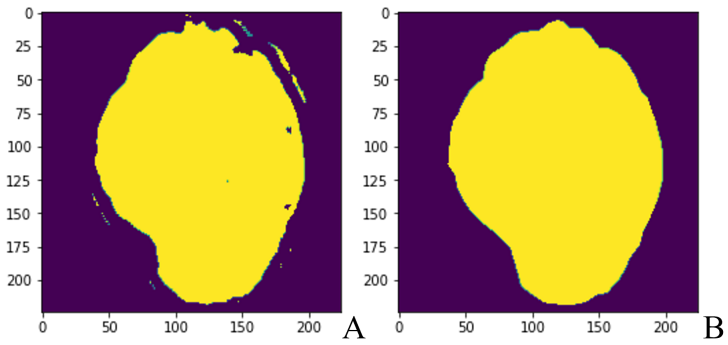


Figura 4-1: Comparación mejores imágenes modelo de 10 épocas. A) Predicción entrenamiento, J = 0,9597 B) Ground Truth

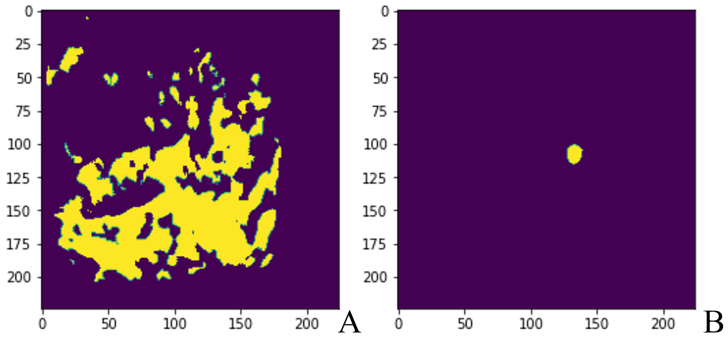


Figura 4-2: Comparación peores imágenes modelo de 10 épocas. A) Predicción test, J = 0,0062 B) Ground Truth

de gran utilidad, pues en el caso de melanoma donde las segmentaciones son muy diferentes entre sí debido al tamaño, su aplicación permite tener una evaluación más imparcial acerca del desempeño de los modelos.

Con lo mencionado anteriormente, se puede observar gráficamente con las figuras **4-1**, **4-2**, **4-3** y **4-4**, que representan la segmentación de los dos modelos con el mejor y peor índice de Jaccard, respectivamente; en el Anexo A se pueden observan más imágenes que presentan las mismas características aquí descritas. Con base en estas figuras, se observa que ambos modelos segmentan mejor aquellas imágenes donde el lunar se encuentra en gran parte de los pixeles, además de presentar un contraste entre el fondo, la piel, y el mismo lunar; esto se sustenta debido a que la base de datos cuenta con algunas imágenes que presentan una especie de filtro azul, como la figura **4-5**, lo que dificulta el rendimiento del modelo para realizar la tarea especificada.

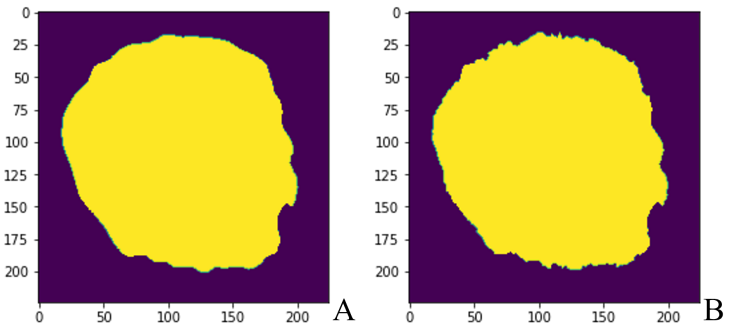


Figura 4-3: Comparación mejores imágenes modelo de 20 épocas. A) Predicción entrenamiento, J = 0,9787 B) Ground Truth

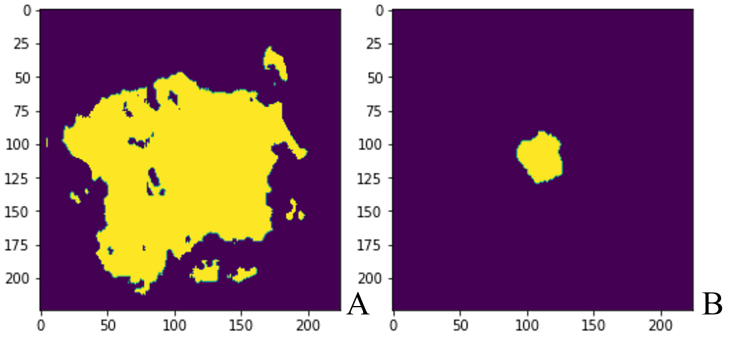


Figura 4-4: Comparación peores imágenes modelo de 20 épocas. A) Predicción entrenamiento, J = 0,0534 B) Ground Truth

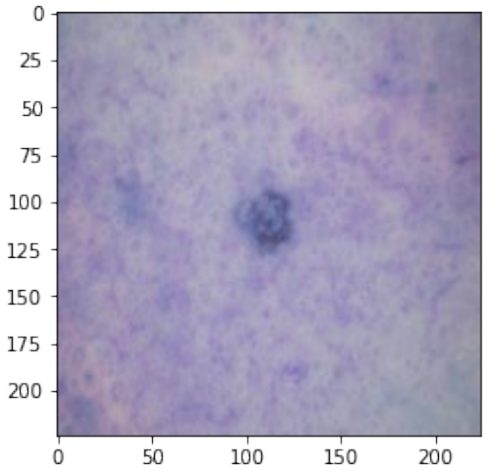


Figura 4-5: Imagen RGB del melanoma de la figura **4-4**

4.2 Clasificación

Durante el desarrollo de la fase de clasificación, se emplearon dos arquitecturas diferentes, y se modificaron varios parámetros para observar el cambio de desempeño de todos los modelos, lo cual se comparó con diferentes métricas que se pueden observar en el Anexo B del presente documento, donde se encuentra una recopilación de las tablas con los resultados para todos los modelos obtenidos. Se observaron diferentes tipos de valores, cada uno con una razón de vital importancia para el determinar el rendimiento a detalle, a continuación, se mostrarán los resultados obtenidos para los diferentes parámetros modificados.

# épocas	Exactitud	Especificidad	Sensibilidad	Kappa
50	0,8078	NAN	0,8078	0
100	0,9111	0,7514	0,9524	0,721
150	0,8333	0,5706	0,8915	0,451
200	0,7	0,372	0,9386	0,335

Tabla 4-2: Métricas de desempeño para entrenamiento de la arquitectura ResNet50 con optimizador Adam 0,0001

Como primer parámetro se varió el valor del optimizador Adam, entre 0,001 y 0,0001, para la arquitectura ResNet50 se notó que para el menor valor las métricas de desempeño tienden a tener mejores resultados, tanto para el conjunto de entrenamiento y test, esto se puede apreciar en la tabla **4-2** y **4-3**, especialmente para el valor de especificidad, que indica la cantidad de verdaderos negativos que el modelo etiqueto como negativos, se observa que el mejor valor obtenido es de 0,7514, correspondiente al conjunto de entrenamiento, lo que significa que aproximadamente el 70% de los casos de la base de datos fueron asignados correctamente a su categoría de 1, es decir, melanoma.

# épocas	Exactitud	Especificidad	Sensibilidad	Kappa
50	0,7611	0,3279	0,829	0,133
100	0,8222	0,6585	0,83	0,193
150	0,7344	0,3796	0,8898	0,3
200	0,6956	0,3407	0,8885	0,256

Tabla 4-3: Métricas de desempeño para entrenamiento de la arquitectura ResNet50 con optimizador Adam 0,0001

Con cada valor del optimizador Adam dentro de la misma arquitectura, se modificaron la cantidad de épocas, para todos los modelos obtenidos se observó que aquellos con 50 épocas,

son los que suelen tener peor desempeño, como se ve en la tabla **4-2**, para este modelo el Kappa tiene un valor de cero puesto que clasifica todas las imágenes como benignas, lo que quiere decir, con base en lo expuesto en la tabla **2-2**, que el nivel de acuerdo entre el experto que asignó las etiquetas y este modelo es leve, o en otras palabras existe poco acuerde entre las partes mencionadas.

Por otro lado, el modelo de 100 épocas presenta unas métricas excelentes, una exactitud de 0,9111 para entrenamiento y 0,8222 para test, lo que significa que la relación entre predicciones correctas con el total de predicciones es bastante elevada, este valor se ve apoyado con los valores que obtiene para las otras métricas.

Con base en lo anterior, es que se selecciona el modelo de 100 épocas para modificar el valor del *dropout* y determinar cómo este afecta el desempeño, al observar la tabla **4-4** se ve una ligera mejora en el rendimiento del modelo especialmente para un *dropout* de 0,75, sin embargo, al probar con el conjunto de prueba, tabla **4-5**, el desempeño baja bastante, presentando errores principalmente para la predicción del melanoma, las razones de esto se explicaran en el capítulo 5.

Dropout	**Exactitud**	**Especificidad**	**Sensibilidad**	**Kappa**
0,25	0,9133	0,8417	0,9264	0,688
0,5	0,93	0,8667	0,9427	0,763
0,75	0,9367	0,8671	0,9515	0,789

Tabla 4-4: Métricas de desempeño para entrenamiento del mejor modelo de ResNet50 con optimizador Adam 0,0001

Dropout	**Exactitud**	**Especificidad**	**Sensibilidad**	**Kappa**
0,25	0,7441	0,1944	0,8017	-0,003
0,5	0,7546	0,25	0,8076	0,039
0,75	0,752	0,2121	0,8035	0,01

Tabla 4-5: Métricas de desempeño para test del mejor modelo de ResNet50 con optimizador Adam 0,0001

Este mismo procedimiento se realizó con la arquitectura InceptionV3, primero se modificó el valor del optimizador Adam por los mismos mencionados anteriormente, y se notó el mismo comportamiento para los conjuntos tanto de entrenamiento como de test donde la exactitud presenta un valor más alto para aquellos modelos en los que se empleó el valor de 0,0001, como se puede observar en las tablas **4-6** y **4-7**, donde el modelo que presento las mejores

métricas para ambos grupos muestra una exactitud de hasta 0,9511 y un valor de especificidad de 0,964.

# épocas	Exactitud	Especificidad	Sensibilidad	Kappa
50	0,8078	NAN	0,8078	0
100	0,8233	0,9375	0,8213	0,13
150	0,8622	0,9455	0,8568	0,401
200	0,9511	0,964	0,9488	0,83

Tabla 4-6: Métricas de desempeño para entrenamiento de la arquitectura InceptionV3 con optimizador Adam 0,0001

# épocas	Exactitud	Especificidad	Sensibilidad	Kappa
50	0,8144	0,7143	0,816	0,08
100	0,8584	0,7654	0,8676	0,435
150	0,8256	0,5755	0,8589	0,341
200	0,82	0,5495	0,858	0,329

Tabla 4-7: Métricas de desempeño para entrenamiento de la arquitectura InceptionV3 con optimizador Adam 0,001

El siguiente parámetro por observar fue la cantidad de épocas bajo los dos optimizadores empleados, se obtuvo un mejor comportamiento para los modelos de 200 y 100 épocas en el conjunto de entrenamiento, como se observa en las tablas **4-6** y **4-7**, sin embargo, se observa una disminución radical del desempeño a la hora de probar todos los modelos en el conjunto de test, alguno de ellos obteniendo un Kappa de valor negativo, lo que quiere decir que no hay ningún acuerdo entre lo establecido por el experto y lo que expone el modelo, en la siguiente sección se expondrá las posibles razones por las que esto pudo suceder con base en lo observado y una revisión literaria.

Dropout	Exactitud	Especificidad	Sensibilidad	Kappa
0,25	0,7388	0,1667	0,7988	-0,023
0,5	0,723	0,1591	0,797	-0,034
0,75	0,7361	0,2093	0,8036	0,01

Tabla 4-8: Métricas de desempeño para test del mejor modelo de InceptionV3 con optimizador Adam 0,0001

Finalmente, con los dos mejores modelos de esta arquitectura se modifico el valor del *dropout*,

y a diferencia de lo que sucedió con los modelos de la ResNet50, el uso de esta variable disminuyo el rendimiento de manera considerable, especialmente para el modelo con Adam de 0,0001 y 200 épocas, donde se observan valores bastante bajos para especificidad, y coeficiente Kappa, esto indica que los modelos presentados en la tabla **4-8** clasifican muy pocas imágenes como melanoma de manera acertada, lo que puede atribuirse a diferentes características de la red neuronal que pueden verse afectadas por usar este parámetro, por ejemplo que esta sea relativamente pequeña en comparación con las imágenes que le entran y sus características.

5 DISCUSIÓN

En el presente capítulo se expone el análisis de resultados obtenidos para las distintas fases del proyecto, así como una comparación entre los diferentes modelos para determinar cómo afecta el cambio de parámetros en su desempeño, y entre las arquitecturas.

Al observar los resultados de segmentación, se nota que aquellas imágenes que presentan similitud entre el fondo, piel, y el lunar (bajo contraste) son las que exhiben más problemas a la hora de ser segmentadas por los modelos, esto se debe a que, si bien segmentar lunares es una tarea ya de por si compleja para las redes convolucionales debido a que la piel no es exactamente igual en todos los pacientes, el hacerlo bajo estas circunstancias se complica aún más; puesto que la extracción de características puede llegar a tomar parte del fondo de la imagen como si fuera el lunar, generando segmentaciones como las mostradas en el capítulo anterior y el Anexo A.

Lo anterior también fue evidente en el estudio realizado por Phan, Kim, Yang y Lee en 2021 [39], donde usan la misma base de datos, y proponen una modificación de la arquitectura U-Net, aplicando módulos de Kernel Selectivo, los cuales evalúan la combinación de información y seleccionan escalas espaciales que resulten ser efectivas para ajustar el campo receptivo de la imagen y permitir el paso de información más grande al encoder, al aplicarla observan que problemas como la variación del tamaño de las lesiones de la piel o, como el mencionado anteriormente, el bajo contraste entre la lesión y el tejido sano no perjudican el rendimiento del modelo, gracias a las técnicas que aplicaron como el Kernel mencionado al inicio de este párrafo, pues reportan un índice de Jaccard de 0,8922, más alto que los reportados por otros estudios mencionados en este mismo artículo [39].

Adicionalmente, en el estudio [40] realizan la misma tarea con la arquitectura U-Net cambiando la cantidad de capas que esta cuenta, en este caso se realiza un post procesamiento para eliminar huecos de la segmentación del modelo y ruido sal-pimienta, en este caso compararon las imágenes antes y después del tratamiento realizado obteniendo como resultado mejores métricas para aquellas imágenes tratadas, especialmente para el índice de Jaccard. Con base en esto, concuerda con la idea general de que la arquitectura U-net resulta ser de gran utilidad para las tareas de segmentación, mostrando un excelente desempeño tanto en datasets pequeño, como el del estudio mencionado, como de tamaño mediano a grande como el empleado en este proyecto; por otro lado, si bien los resultados obtenidos son muy fa-

vorables en comparación con la segmentación del experto, una etapa de post procesamiento puede llegar a mejorar algunas máscaras binarias obtenidas por del modelo todo esto teniendo en cuenta que la selección correcta de la región más grande afecta en gran medida el desempeño del modelo [40].

Al realizar una comparación entre los resultados obtenidos en este proyecto y los obtenidos por otros investigadores, existen puntos de similitud como los problemas de segmentación que presentan las imágenes con bajo contraste, con la diferencia que en el estudio realizado por Phan et al. [39], el uso del Kernel selectivo permitió que el desempeño del modelo empleado no se viera perjudicado. Por otro lado, los resultados de Araujo en 2020, [40] si presentan una diferencia en la segmentación, pues allí realizaron una etapa de post-procesamiento para eliminar cualquier tipo de artefacto que estuviera presente en las imágenes, esto demostró ser de gran utilidad, pues el índice de Jaccard alcanzó un valor de 0,873 en comparación con el mejor valor obtenido para test en este proyecto, 0,7796, como se ve en la tabla **4-1**.

Otro aspecto para tener en cuenta en esta fase y en la siguiente fue el uso de diferente número de épocas, de manera general se observó que el desempeño del modelo, mejoraba cuantas más épocas presentaba, por ejemplo, en la segmentación, el de 20 épocas tuvo mejores índices de Jaccard y Dice que el de 10 épocas. Esto se debe a que la red convolucional ajustaba mejor sus pesos cuando el entrenamiento tenia mayor cantidad de épocas, obteniendo mejores características y por ende un mejor desempeño en clasificación. Sin embargo, puede ocurrir el caso que ocurrió en la fase de clasificación, en donde el conjunto de entrenamiento obtiene excelentes métricas de desempeño, pero al probar con el conjunto de prueba este disminuía considerablemente, es decir, se presenta un problema de overfitting.

En la fase de clasificación, se modificaron los valores para el optimizador Adam y el dropout, para el primero de estos se encontró que aquellos con menor valor obtienen mejores métricas de desempeño en entrenamiento, mientras que en test este desempeño disminuye, esto se le atribuye a que la red neuronal al usar una tasa de aprendizaje menor aprende más lento al inicio generando que pueda existir un problema de *overfitting*, este comportamiento fue similar para el otro valor de Adam, con la diferencia que en entrenamiento sus resultados no destacaron tanto. Con base en estos resultados se modificó el valor del dropout, el cual tiene por objetivo ocultar imágenes durante el entrenamiento de la red neuronal para evitar el problema mencionado anteriormente, al emplearlo mejoro el rendimiento en algunas ocasiones, sin embargo, no se observó un cambio significativo que indicara que disminuyo el overfitting, pues la disminución de desempeño entre ambos conjuntos se mantuvo constante.

Ahora bien, de manera general para las dos arquitecturas empleadas, los resultados mostraron una mejor clasificación de melanoma para la ResNet50 en el conjunto de entrenamiento, pero InceptionV3 muestra, en comparación, mejores valores de especificidad en test en com-

paración. Algo similar ocurre en un estudio realizado en 2020 [41], donde emplean diferentes arquitecturas para la clasificación de diferentes objetos de la base de datos ImageNet2015, reportan que la Resnet50 muestra un buen desempeño en términos de exactitud, pero se evidencia la persistencia del problema de overfitting, este disminuye con el uso de InceptionV3 pero resaltan que los resultados de entrenamiento son mucho peores que los otros modelos empleados.

La revisión sistemática realizada en el 2021 por Höhn et al. se enfoca en el impacto que tiene el uso de este tipo de herramientas de detección de cáncer de piel, como criterios de inclusión tuvieron en cuenta la cantidad de pacientes empleados, algoritmos basados en las redes neuronales convolucionales y si hacían clasificación binaria (melanoma y no melanoma) o multiclase (entre los diferentes tipos de cáncer de piel); encontraron que el rendimiento para los 11 estudios encontrados es bueno, con un promedio de exactitud mayor a 0.6, sin embargo, el desempeño de estos aumentaba cuando además de las imágenes se contaba con información del paciente como edad, género, ubicación del lunar, tamaño, entre otros [42]. Con base en esta información y el análisis realizado, se entiende que el impacto clínico que tienen este tipo de herramientas es bastante elevado, pues su uso tiene la posibilidad de disminuir los tiempos de diagnóstico, resultando ser aún más confiable cuando se cuenta con información adicional y no solamente con la imagen del dermatológica.

Finalmente, la comparación de los resultados y la literatura mencionada exponen que la arquitectura ResNet50 presenta un buen desempeño para tareas de clasificación, pero en el estudio de Li et al. mencionan que presenta overfitting [41], problema que, si se presentó para algunos modelos de este proyecto, pero no de forma general. Para la otra arquitectura, el mismo estudio menciona que InceptionV3 no tiene un buen desempeño para clasificar, al igual que en el estudio realizado por Cassidy et al., donde para las mismas imágenes de ISIC, esta arquitectura presentó uno de los peores valores de exactitud, 0,55 [33]; con la diferencia el problema de overfitting disminuía con el uso de esta arquitectura, sin embargo, los resultados obtenidos en este proyecto evidencian más persistentemente este inconveniente para sus diferentes modelos que para los obtenidos con ResNet50.

6 CONCLUSIONES

El aprendizaje profundo ofrece herramientas que ayudan a la detección automática de melanoma para el procesamiento y análisis de imágenes dermatológicas. Con los resultados obtenidos en segmentación y clasificación de melanomas, consideramos que los modelos desarrollados en este trabajo dirigido podrían ser una herramienta potencialmente útil que puede permitir la disminución del tiempo de detección y diagnóstico del melanoma.

Por otro lado, se logró establecer la importancia que tienen distintos parámetros en la definición de los modelos por medio de una comparación entre los resultados obtenidos para el conjunto de entrenamiento y test. Allí se determino que el número de épocas, el optimizador y el dropout influyen en gran medida en el desempeño del modelo, en la presencia o no de overfitting, lo que genera una disminución en el rendimiento del modelo a la hora de mirar los resultados obtenidos en test. Lo anterior, se encuentra reportado en diversos estudios que emplean metodologías similares, bien sea para imágenes de melanoma u otro tipo de imágenes.

En contraparte, resulta evidente con la observación de las imágenes de la base de datos, y con base en los resultados aquí expuestos, que el desarrollo de herramientas de diagnóstico puede resultar de gran importancia para usuarios no médicos que quieran estar pendientes de esta enfermedad debido a distintos factores de riesgo que tienen y para médicos generales, los cuales debido a que generalmente no están expuestos a este tipo de enfermedades y por lo cual no tienen el conocimiento suficiente para detectarla rápida y eficientemente, pueden orientar en mejor medida a sus pacientes para una opinión dermatológica experta.

Es por esto, que lo desarrollado en este proyecto es un gran punto de partida para el desarrollo de una herramienta de apoyo diagnóstico que permita a los potenciales usuarios obtener una idea acerca de las características del lunar y de esta forma disminuir la mortalidad que presenta esta neoplasia hoy en día. Para ello la segmentación, clasificación y uso de la regla ABCD son vitales para el desarrollo de este proyecto, pues a partir de ellas se obtienen las características del lunar que permiten determinar el riesgo que este presenta de ser o convertirse en melanoma.

Finalmente, es importante tener en cuenta que el uso de diferentes modelos y arquitecturas son un factor a tener presente y su selección debe hacerse partiendo de la base de los objetivos

que se desean, por ejemplo, para evitar overfitting algunas arquitecturas lo logran mejor que otras, así mismo ocurre con las imágenes de la base de datos que se este trabajando, ya que, una imagen que contenga mucho ruido o tenga un contraste bajo entre la piel y el lunar afectan también el desempeño del modelo.

7 TRABAJOS FUTUROS

Para la continuación de este proyecto, con el fin de seguir contribuyendo al desarrollo de una herramienta de apoyo diagnóstico, se propone el uso o desarrollo de una base de datos mayor que cuente con un protocolo más estricto para la toma de las imágenes de forma que estas cuenten con las mismas condiciones como el nivel de luz, el mínimo ruido posible entre otras.

De igual forma, a mediano plazo el uso de sistemas de detección como el YOLO, para que, con lo obtenido en este proyecto, se puedan detectar los lunares, dar sus características y el posible riesgo que este sea melanoma, esto con el fin de que, a largo plazo, se pueda realizar la detección en tiempo real de estos y con esto desarrollar una aplicación que permita a los usuarios ingresar los factores de riesgo que presenten, tomarle foto a los lunares para guardar esta información y evidenciar algún cambio morfológico que pueda ser un signo de alerta.

Finalmente, otro trabajo a futuro es el obtener la opinión de dermatólogos expertos en la detección de este cáncer de piel que nos brinden mayor información acerca de otras señales de alerta visibles para agregarlas y tenerlas en cuenta para el apoyo diagnóstico que permita disminuir el nivel de mortalidad que tiene el melanoma hoy en día.

8 ANEXO A

Se presentan las mejores y peores segmentaciones de los dos modelos implementados para esta parte del proyecto, se presentan en orden de mayor a menor según el índice de Jaccard.

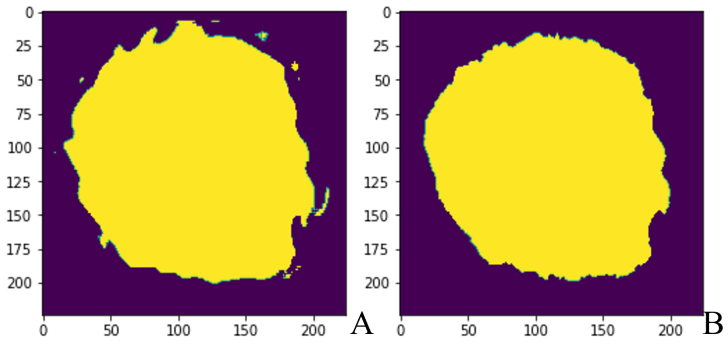


Figura 8-1: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,9453 B) Ground Truth

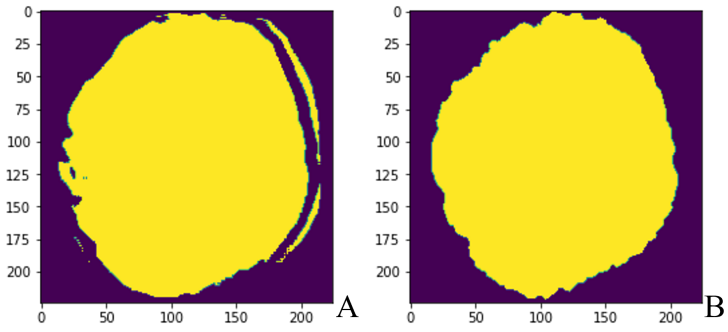


Figura 8-2: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,9406 B) Ground Truth

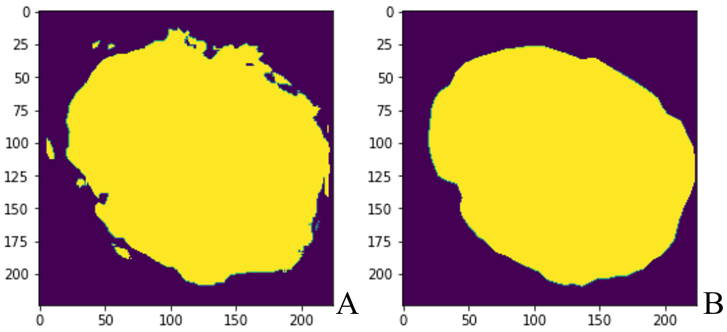


Figura 8-3: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,9253 B) Ground Truth

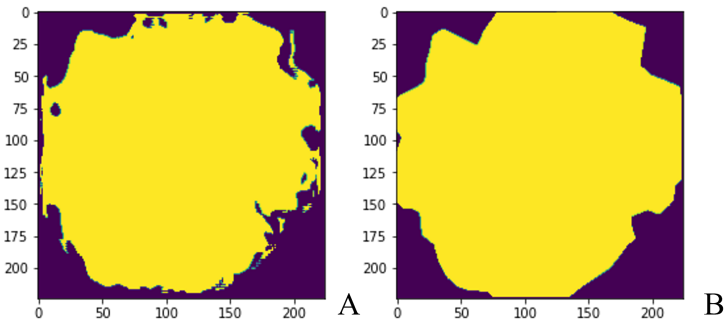


Figura 8-4: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción en entrenamiento, J = 09235, B) Ground Truth

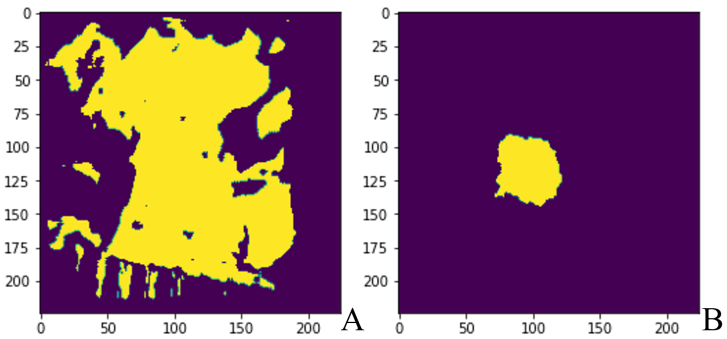


Figura 8-5: Comparación peores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,0920 B) Ground Truth

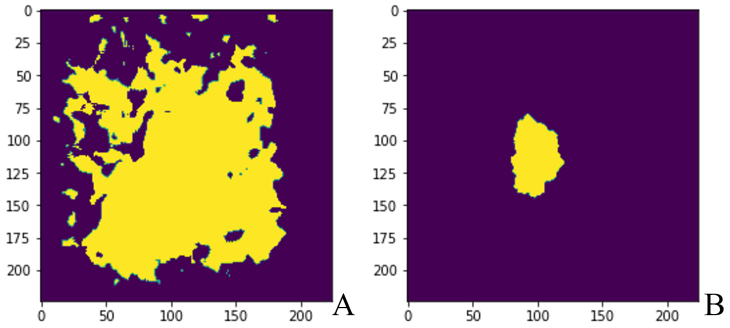


Figura 8-6: Comparación peores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,0841 B) Ground Truth

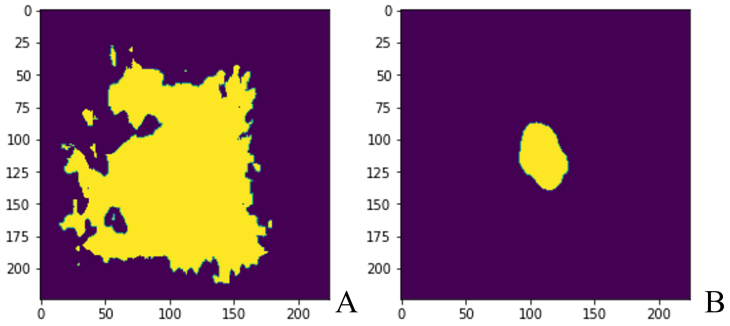


Figura 8-7: Comparación peores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,0817 B) Ground Truth

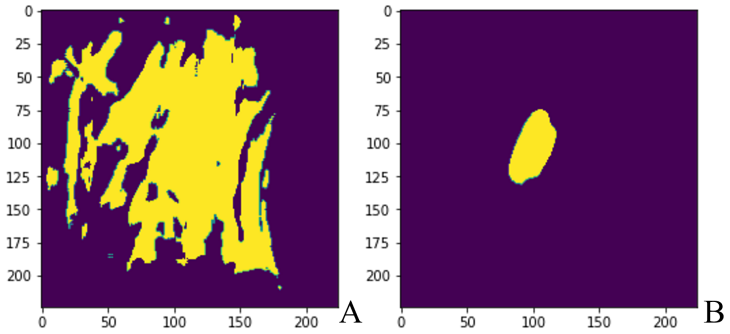


Figura 8-8: Comparación peores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,0735 B) Ground Truth

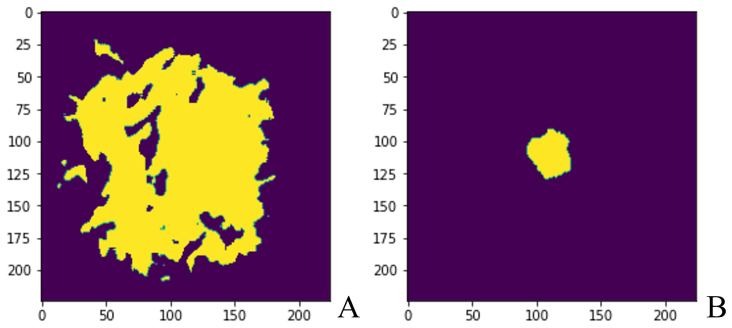


Figura 8-9: Comparación peores imágenes modelo de 10 épocas.
A) Predicción entrenamiento, J = 0,0515 B) Ground Truth

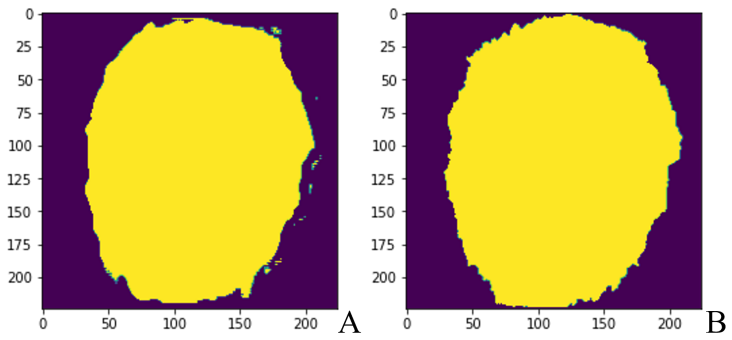


Figura 8-10: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,9472 B) Ground Truth

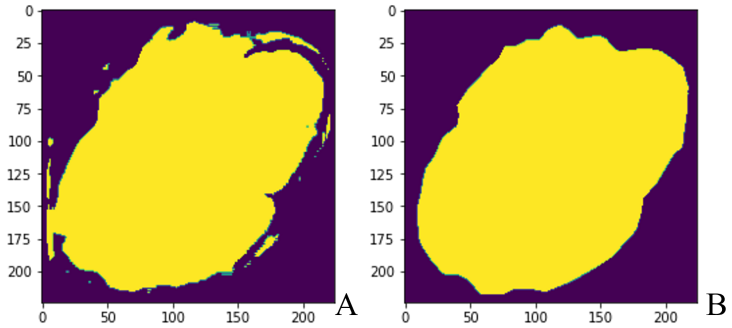


Figura 8-11: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,9199 B) Ground Truth

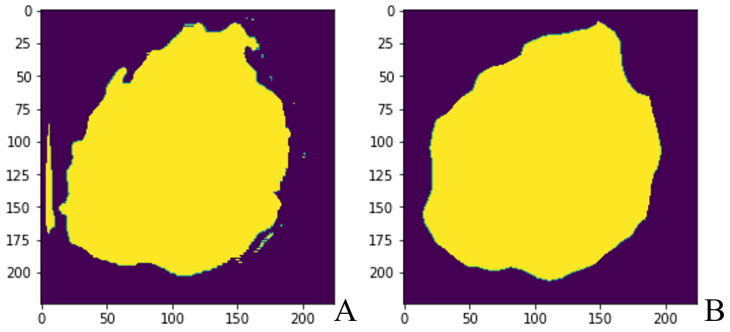


Figura 8-12: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,9046 B) Ground Truth

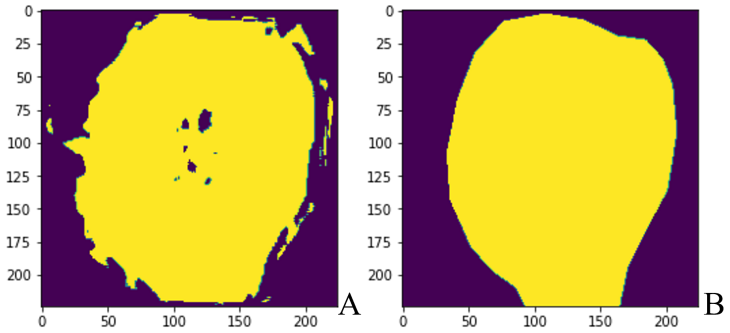


Figura 8-13: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,9005 B) Ground Truth

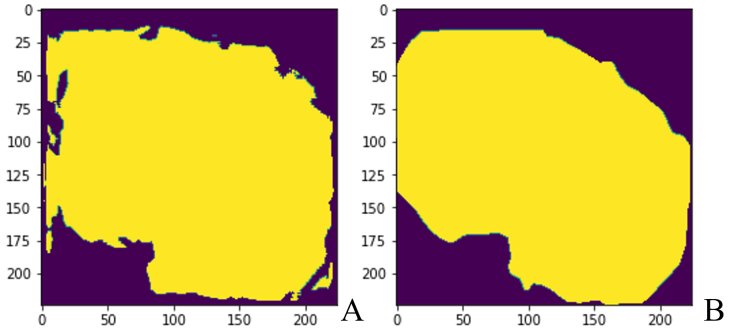


Figura 8-14: Comparación mejores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,89966 B) Ground Truth

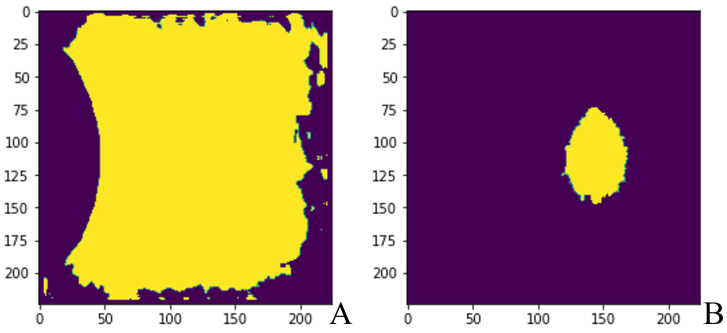


Figura 8-15: Comparación peores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,0716 B) Ground Truth

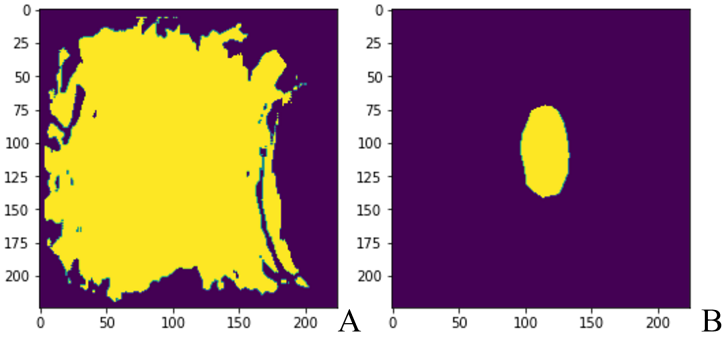


Figura 8-16: Comparación peores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,0664 B) Ground Truth

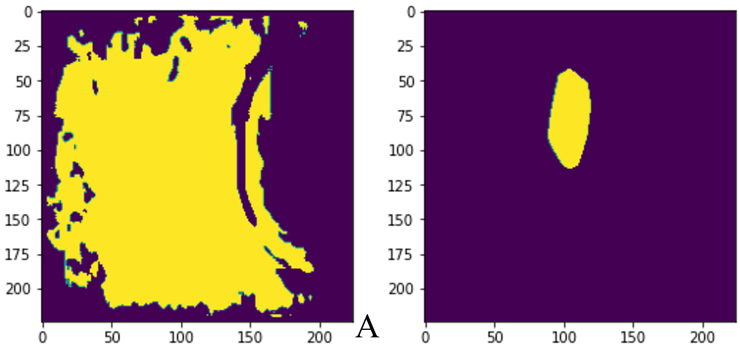


Figura 8-17: Comparación peores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,0595 B) Ground Truth

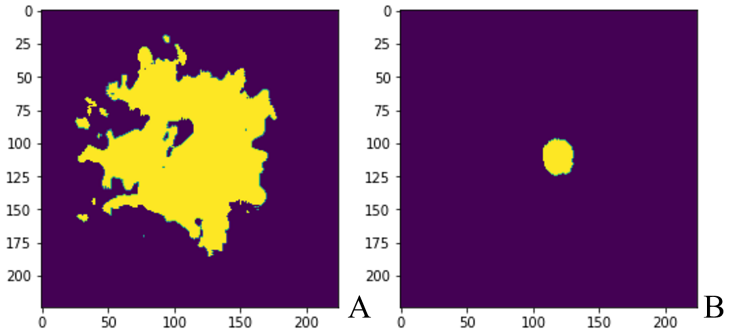


Figura 8-18: Comparación peores imágenes modelo de 10 épocas.
A) Predicción test, J = 0,0417 B) Ground Truth

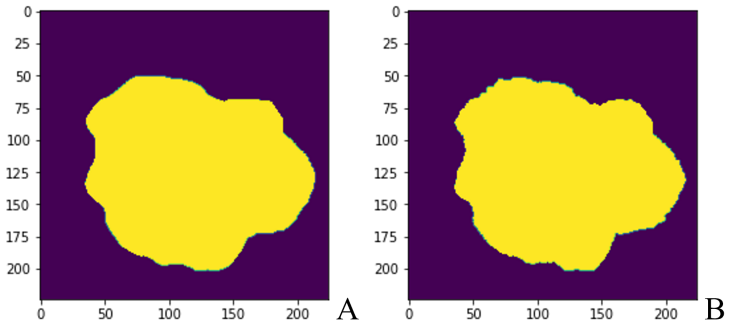


Figura 8-19: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,9750 B) Ground Truth

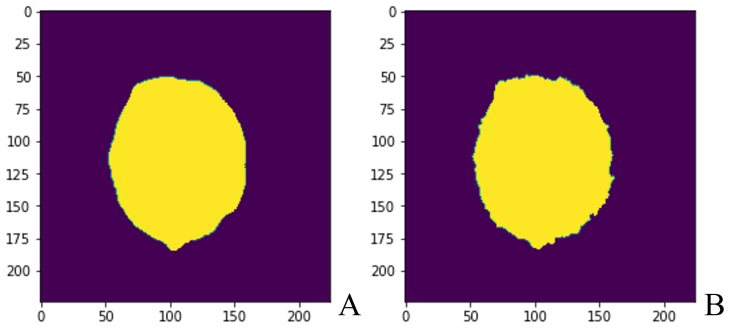


Figura 8-20: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,9724 B) Ground Truth

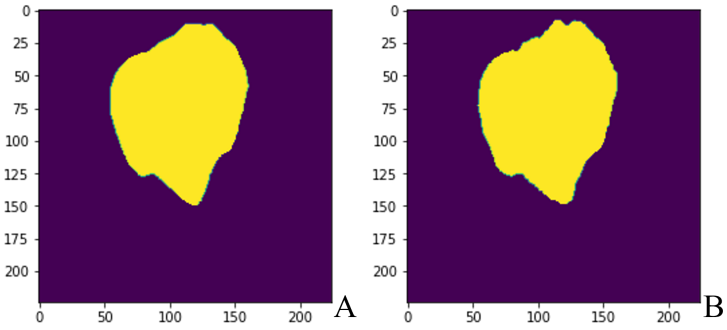


Figura 8-21: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,9711 B) Ground Truth

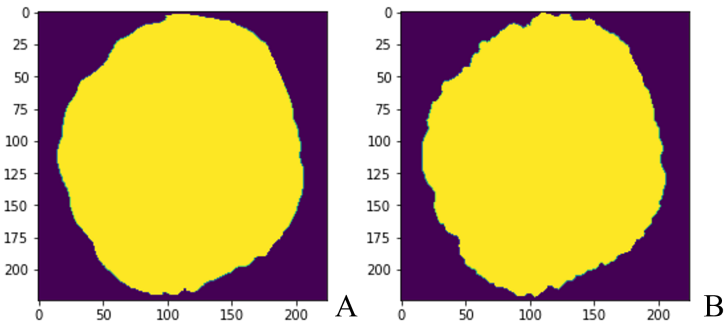


Figura 8-22: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,9708 B) Ground Truth

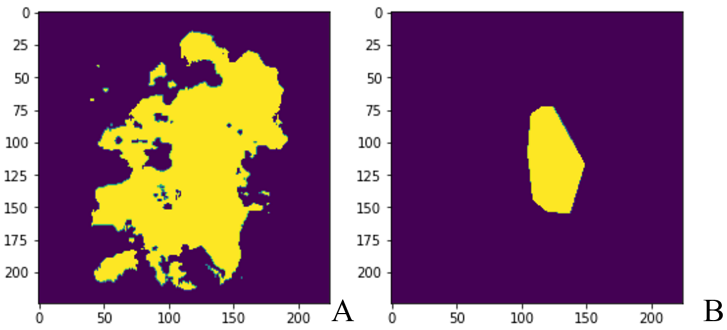


Figura 8-23: Comparación peores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,1671 B) Ground Truth

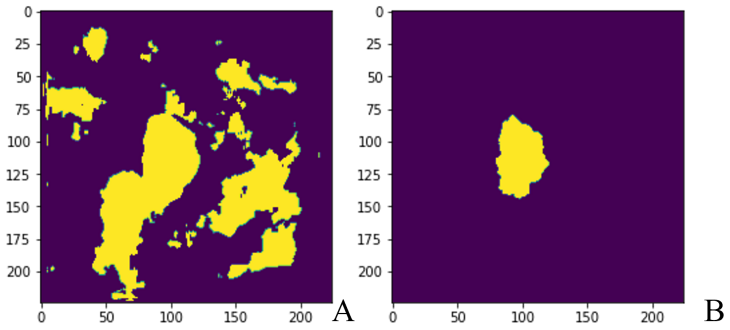


Figura 8-24: Comparación peores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,1503 B) Ground Truth

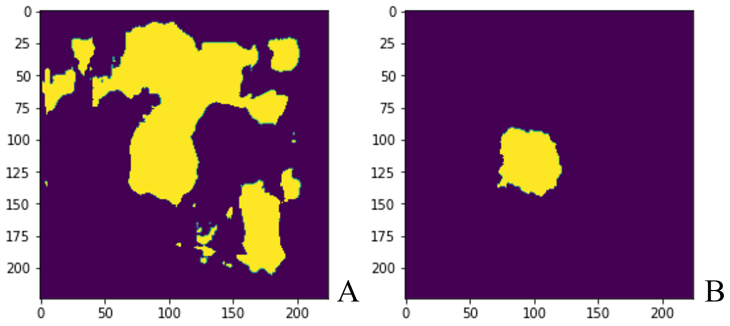


Figura 8-25: Comparación peores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,1496 B) Ground Truth

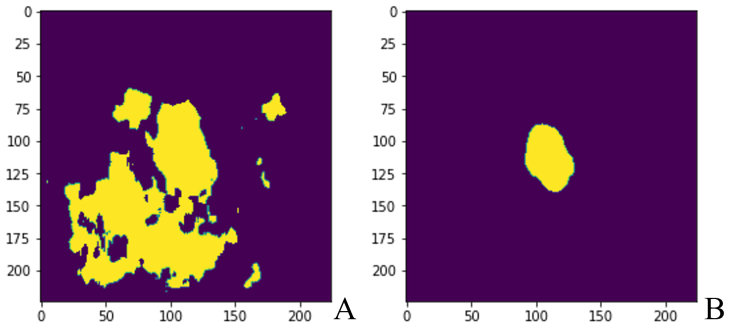


Figura 8-26: Comparación peores imágenes modelo de 20 épocas.
A) Predicción entrenamiento, J = 0,1349 B) Ground Truth

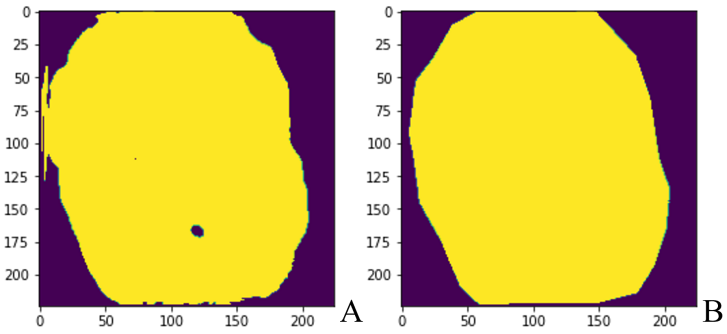


Figura 8-27: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,9689 B) Ground Truth

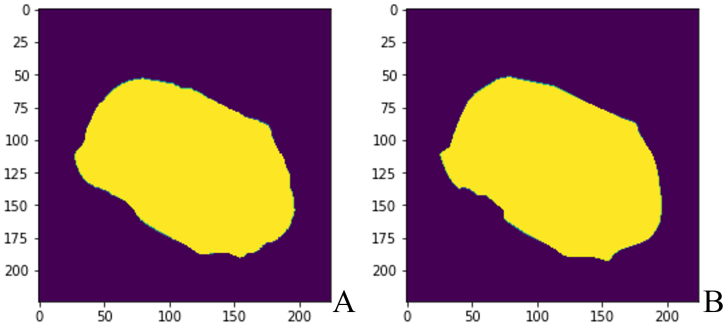


Figura 8-28: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,9679 B) Ground Truth

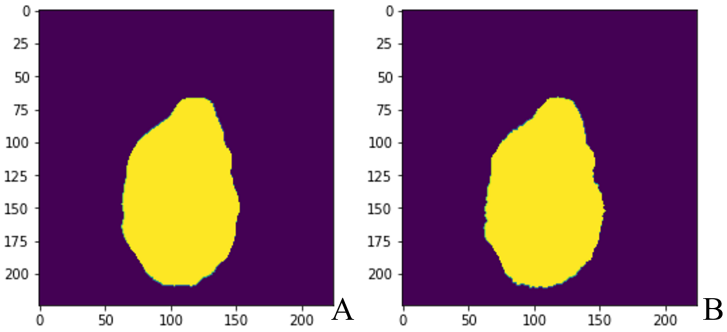


Figura 8-29: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,9672 B) Ground Truth

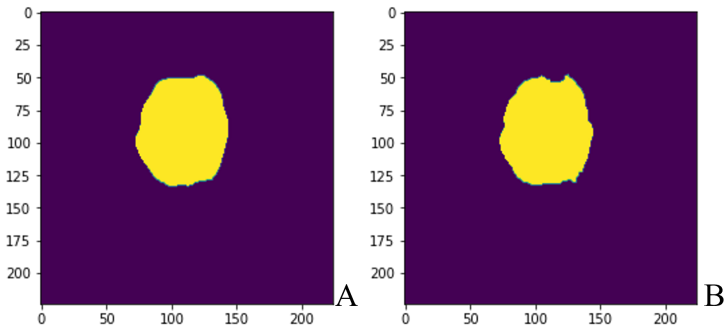


Figura 8-30: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,9623 B) Ground Truth

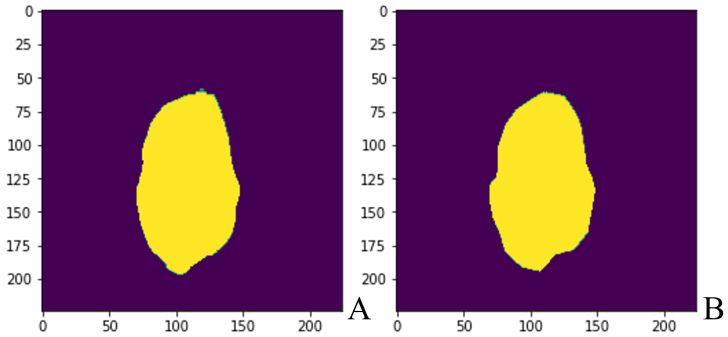


Figura 8-31: Comparación mejores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,9600 B) Ground Truth

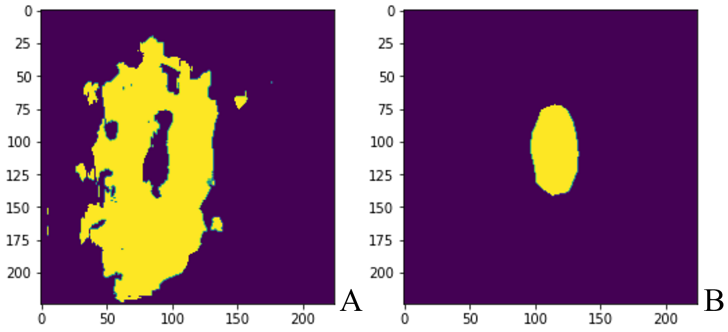


Figura 8-32: Comparación peores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,1474 B) Ground Truth

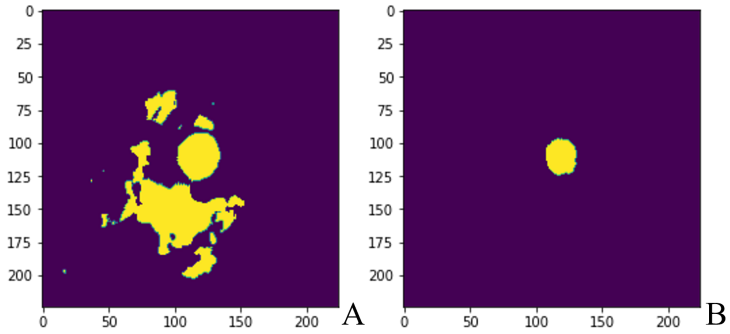


Figura 8-33: Comparación peores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,1126 B) Ground Truth

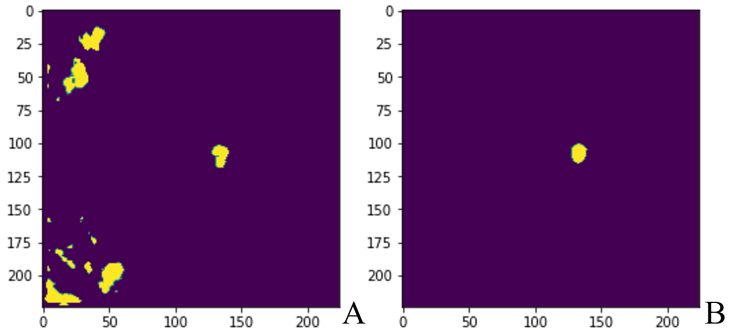


Figura 8-34: Comparación peores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,0830 B) Ground Truth

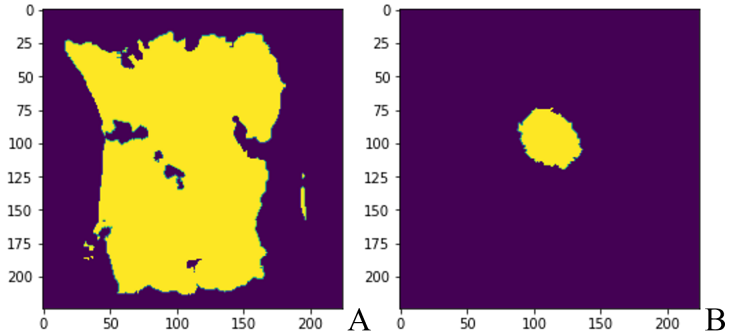


Figura 8-35: Comparación peores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,0657 B) Ground Truth

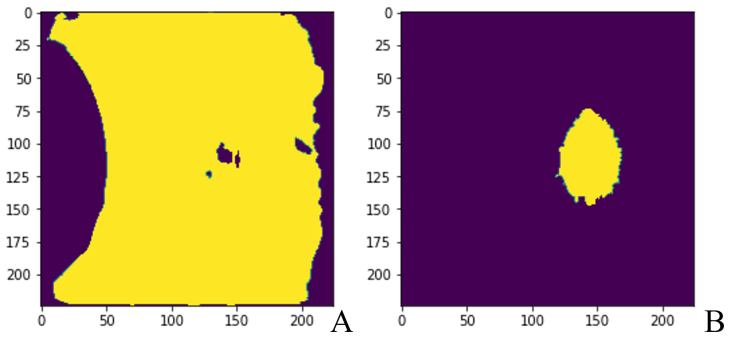


Figura 8-36: Comparación peores imágenes modelo de 20 épocas.
A) Predicción test, J = 0,0595 B) Ground Truth

9 ANEXO B

A continuación se muestran las tablas correspondientes a las métricas de desempeño obtenidas para la clasificación de las imágenes dermatológicas, se presenta para la arquitectura ResNet50 e InceptionV3

# épocas	Accuracy	Especificidad	Sensitividad	Kappa
50	0,8021	NAN	0,8021	0
100	0,7071	0,2857	0,8271	0,117
150	0,686	0,1944	0,8013	-0,004
200	0,5172	0,2	0,804	0,004

Tabla 9-1: Métricas de desempeño para test de la arquitectura ResNet50 con optimizador Adam 0,0001

# épocas	Accuracy	Especificidad	Sensitividad	Kappa
50	0,7414	0,2444	0,8084	0,041
100	0,7678	0,2593	0,8068	0,036
150	0,6095	0,1712	0,791	-0,042
200	0,628	0,225	0,8147	0,044

Tabla 9-2: Métricas de desempeño para test de la arquitectura ResNet50 con optimizador Adam 0,001

Dropout	Accuracy	Especificidad	Sensibilidad	Kappa
0,25	0,8078	NAN	0,8078	0
0,5	0,798	0,4194	0,8258	0,133
0,75	0,8178	0,5495	0,848	0,284

Tabla 9-3: Métricas de desempeño para entrenamiento del mejor modelo de ResNet50 con optimizador Adam 0,001

Dropout	Accuracy	Especificidad	Sensibilidad	Kappa
0,25	0,8021	NAN	0,8021	0
0,5	0,752	0,1935	0,8017	-0,003
0,75	0,7414	0,2195	0,8047	0,018

Tabla 9-4: Métricas de desempeño para test del mejor modelo de ResNet50 con optimizador Adam 0,001

# épocas	Accuracy	Especificidad	Sensibilidad	Kappa
50	0,8021	NAN	0,8021	0
100	0,8021	0,5	0,8053	0,031
150	0,7889	0,3077	0,806	0,034
200	0,7836	0,36	0,8136	0,09

Tabla 9-5: Métricas de desempeño para test de la arquitectura InceptionV3 con optimizador Adam 0,0001

Dropout	Accuracy	Especificidad	Sensibilidad	Kappa
0,25	0,8678	0,7647	0,881	0,495
0,5	0,91	0,877	0,9152	0,674
0,75	0,92	0,8797	0,927	0,718

Tabla 9-6: Métricas de desempeño para entrenamiento del mejor modelo de InceptionV3 con optimizador Adam 0,0001

# épocas	Accuracy	Especificidad	Sensibilidad	Kappa
50	0,7968	0,3333	0,8043	0,021
100	0,7731	0,2963	0,8097	0,058
150	0,7203	0,1837	0,8	-0,013
200	0,7018	0,12	0,7903	-0,074

Tabla 9-7: Métricas de desempeño para test de la arquitectura InceptionV3 con optimizador Adam 0,001

Dropout	Accuracy	Especificidad	Sensibilidad	Kappa
0,25	0,8189	0,625	0,8279	0,175
0,5	0,3867	0,2116	0,8601	0,044
0,75	0,8267	0,5714	0,8656	0,367

Tabla 9-8: Métricas de desempeño para entrenamiento del mejor modelo de InceptionV3 con optimizador Adam 0,001

Dropout	Accuracy	Especificidad	Sensibilidad	Kappa
0,25	0,7863	0,2857	0,8055	0,029
0,5	0,3905	0,2153	0,8476	0,04
0,75	0,7784	0,3784	0,8216	0,137

Tabla 9-9: Métricas de desempeño para test del mejor modelo de InceptionV3 con optimizador Adam 0,001

Bibliography

[1] Abcde of skin cancer, 5 2019. https://www.conehealth.com/services/cancer-care/skin/abcde-of-skin-cancer/.

[2] Saeed Alzahrani, Waleed Al-Nuaimy, and Baidaa Al-Bander. Seven-point checklist with convolutional neural networks for melanoma diagnosis. pages 211–216. IEEE, 10 2019.

[3] Vahid. Deep learning and its features, 2020. https://cilix.ir/2020/06/04/deep-learning-and-its-features/.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 5 2015.

[5] A. Sai Bharadwaj Reddy and D. Sujitha Juliet. Transfer learning with resnet-50 for malaria cell-image classification. pages 0945–0949. IEEE, 4 2019.

[6] Xiao Tian, Hugh Daigle, and Han Jiang. Feature detection for digital images using machine learning algorithms and image processing. American Association of Petroleum Geologists, 2018.

[7] Fernando Ferreira Lima dos Santos, Lucas Arthur de Almeida Telles, Jorge Tadeu Fim Rosas, Amanda Pereira Assis Gomes, Rodrigo Nogueira Martins, Amélia Laísy do Nascimento, and Emanoel Di Tarso dos Santos Sousa. Open source iterative bayesian classifier algorithm for quality assessment of processed coffee beans. *Nativa*, 8:118, 2 2020.

[8] Liga Colombiana Contra el Cáncer. Datos — cáncer de piel. https://www.ligacancercolombia.org/educacion/datos-cancer-de-piel/.

[9] Flavia Carolina Pozzobon, Álvaro Enrique Acosta, and Juan Sebastián Castillo. Cáncer de piel en colombia: cifras del instituto nacional de cancerología. *Revista de la Asociación Colombiana de Dermatología y Cirugía Dermatológica*, 26:12–17, 4 2018.

[10] Paola Velasquez. Cáncer de piel ¿cómo esta colombia en esta materia?, 6 2020.

[11] Fondo Colombiano de Enfermedades de Alto Costo. Situación del cáncer en la población adulta atendida en el sgsss de colombia 2020. pages 237–264, 2021.

[12] Melanoma Research Alliance. Melanoma statistics. https://www.curemelanoma.org/about-melanoma/melanoma-statistics-2/.

[13] H Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3:159–165, 3 2002.

[14] A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown. Projections of the cost of cancer care in the united states: 2010-2020. *JNCI Journal of the National Cancer Institute*, 103:117–128, 1 2011.

[15] Aylen Vanessa Ospina Serrano and Laura Bernal. Tratamiento del melanoma maligno en colombia : resultados encuesta acho y proyecto grupo colombo-español del tratamiento multidisciplinario del melanoma. *Revista Colombiana de Hematología y Oncología*, 6:33–37, 5 2019.

[16] Mary S. Brady, Susan A. Oliveria, Paul J. Christos, Marianne Berwick, Daniel G. Coit, Jared Katz, and Allan C. Halpern. Patterns of detection in patients with cutaneous melanoma. *Cancer*, 89:342–347, 7 2000.

[17] Timo Partanen, Patricia Monge, and Catharina Wesseling. Causas y prevención del cáncer ocupacional. *Acta Médica Costarricense*, 51:195–205, 12 2009.

[18] Thomas Kornek and Matthias Augustin. Skin cancer prevention. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 11:283–298, 4 2013.

[19] Hospital Universitario Centro Dermatológico Federico Lleras Acosta. Tumores, 2015.

[20] Miguel A. Linares, Alan Zakaria, and Parminder Nizran. Skin cancer. *Primary Care: Clinics in Office Practice*, 42:645–659, 12 2015.

[21] Esteban Uribe, Ángela Londoño, Guillermo Jiménez, Álvaro Sanabria, and Milton Mejía. Carcinoma escamocelular de la piel de alto riesgo: definición, diagnóstico y manejo. *Med Cutan Iber Lat Am*, 45:8–13, 2017.

[22] Carlos A Wilches, Óscar J Perdomo, and César A Perdomo. A method to detect potentially malignant skin lesions through image segmentation. In *World Congress on Medical Physics and Biomedical Engineering 2018*, pages 289–293. Springer, 2019.

[23] Álvaro Enrique Acosta, Eduardo Fierro, Victoria Eugenia Velásquez, and Xavier Rueda. Melanoma: patogénesis, clínica e histopatología. *Revista de la Asociación Colombiana de Dermatología y Cirugía Dermatológica*, 17:87–108, 2 2019.

[24] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B. Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy. *Journal of the American Academy of Dermatology*, 30:551–559, 4 1994.

[25] Pedro Zaballos, Cristina Carrera, Susana Puig, and Josep Malvehy. Criterios dermatoscópicos para el diagnóstico del melanoma. *Medicina cutánea ibero-latino-americana*, 32:3–17, 1 2004.

[26] Lorena Valenzuela De Blasis. Diagnostico temprano de melanoma: Revisión del abcd. *Medicina Familiar*, 2005.

[27] Stefania Borsari and Caterina Longo. Melanoma staging, 2017.

[28] Therapy of melanoma, 2017.

[29] Iván Montenegro, Aleidys Hernandez, Diego Chavarro, María Vélez, Gabo Tovar, Angela Niño, and Alejandro Olaya. Macrotendencias hacia el 2030 el mundo y américa latina. *Colciencias*, 2, 2018.

[30] Deepti Malik, Saniya Mahendiratta, Harpinder Kaur, and Bikash Medhi. Futuristic approach to cancer treatment. *Gene*, 805:145906, 12 2021.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning (Adaptive Computation and Machine Learning series)*. MIT Press, 2016.

[32] Qaiser Abbas, Farheen Ramzan, and Muhammad Usman Ghani. Acral melanoma detection using dermoscopic images and convolutional neural networks. *Visual Computing for Industry, Biomedicine, and Art*, 4:25, 12 2021.

[33] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305, 1 2022.

[34] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). 5 2016.

[35] Jeffrey Girard. Using cohen's kappa statistic for evaluating a binary classifier. https://stats.stackexchange.com/questions/265445/using-cohens-kappa-statistic-for-evaluating-a-binary-classifier.

[36] Isic. Isic challenge, 2016. https://www.isic-archive.com/!/topWithHeader/tightContentTop/about/aboutIsicOverview.

[37] Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. 3 2021.

[38] Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. 3 2021.

[39] Tran-Dac-Thinh Phan, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Skin lesion segmentation by u-net with adaptive skip connection and structural awareness. *Applied Sciences*, 11:4528, 5 2021.

[40] Rafael Luz Araujo, Ricardo de Andrade L. Rabelo, Joel J. P. C. Rodrigues, and Romuere R. V. e Silva. Automatic segmentation of melanoma skin cancer using deep learning. pages 1–6. IEEE, 3 2021.

[41] Wenmei Li, Ziteng Wang, Yu Wang, Jiaqi Wu, Juan Wang, Yan Jia, and Guan Gui. Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1986–1995, 2020.

[42] Julia Höhn, Achim Hekler, Eva Krieghoff-Henning, Jakob Nikolas Kather, Jochen Sven Utikal, Friedegund Meier, Frank Friedrich Gellrich, Axel Hauschild, Lars French, Justin Gabriel Schlager, Kamran Ghoreschi, Tabea Wilhelm, Heinz Kutzner, Markus Heppt, Sebastian Haferkamp, Wiebke Sondermann, Dirk Schadendorf, Bastian Schilling, Roman C Maron, Max Schmitt, Tanja Jutzi, Stefan Fröhling, Daniel B Lipka, and Titus Josef Brinker. Integrating patient data into skin cancer classification using convolutional neural networks: Systematic review. *Journal of Medical Internet Research*, 23:e20708, 7 2021.