DeepL

Suscríbete a DeepL Pro para poder traducir archivos de mayor tamaño.
Más información disponible en www.DeepL.com/pro.

Ingeniería biomédica
En línea

Díaz-Pinto *et al. BioMed Eng OnLine*(2019) 18:29
https://doi.org/10.1186/s12938-019-0649-y

INVESTIGACIÓNAcceso abierto

CNNs para la evaluación automática del glaucoma utilizando imágenes del fondo de ojo: una validación extensiva

Andrés Díaz-Pinto[1]*, Sandra Morales[1], Valery Naranjo[1], Thomas Köhler[2], José M. Mossi[3] y Amparo Navea[4]

*Correspondencia: andiapin@upv.es; vnaranjo@dcom.upv.es
[1] Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, España
La lista completa de información sobre los autores está disponible al final del artículo

Resumen

Antecedentes: La mayoría de los algoritmos actuales para la evaluación automática del glaucoma a partir de imágenes fundus se basan en características elaboradas a mano y basadas en la segmentación, que se ven afectadas por el rendimiento del método de segmentación elegido y las características extraídas.
Entre otras características, las redes neuronales convolucionales (CNN) son conocidas por su capacidad de aprender características altamente discriminativas a partir de las intensidades brutas de los píxeles.

Métodos: En este trabajo, empleamos cinco modelos diferentes entrenados en ImageNet (VGG16, VGG19, InceptionV3, ResNet50 y Xception) para la evaluación automática del glaucoma utilizando imágenes de fondo de ojo. Los resultados de una extensa validación mediante estrategias de validación cruzada y pruebas cruzadas se compararon con trabajos anteriores de la literatura.

Resultados: Utilizando cinco bases de datos públicas (1707 imágenes), se obtuvo un AUC medio de *0,9605* con un intervalo de confianza del 95% de 95,92-97,07%, una especificidad media de *0,8580* y una sensibilidad media de *0,9346* tras utilizar la arquitectura Xception, mejorando significativamente el rendimiento de otros trabajos del estado del arte. Además, se ha hecho pública una nueva base de datos clínicos, ACRIMA, que contiene 705
imágenes etiquetadas. Se compone de 396 imágenes glaucomatosas y 309 imágenes normales, es decir, la mayor base de datos pública para el diagnóstico del glaucoma. La alta especificidad y sensibilidad obtenidas del enfoque propuesto están respaldadas por una extensa validación que utiliza no sólo la estrategia de validación cruzada, sino también la validación de pruebas cruzadas en, según el conocimiento de los autores, todas las bases de datos etiquetadas de glaucoma disponibles públicamente.

Conclusiones: Estos resultados sugieren que el uso de modelos entrenados en ImageNet es una alternativa robusta para el sistema de detección automática del glaucoma. Todas las imágenes, los pesos de las CNN y el software utilizado para afinar y probar las cinco CNN están a disposición del público, lo que podría utilizarse como banco de pruebas para nuevas comparaciones.

Palabras clave: Glaucoma, base de datos ACRIMA, imágenes de fondo de ojo, CNN, ajuste fino

Introducción

El glaucoma es una enfermedad ocular neurodegenerativa irreversible que se considera uno de los principales motivos de discapacidad visual en el mundo [1]. Según la Organización Mundial de la Salud (OMS), el glaucoma afecta a más de 65 millones de personas en todo el mundo [2]. Como puede ser asintomático, la detección y el tratamiento tempranos son importantes para prevenir la pérdida de visión. Esta enfermedad ocular silenciosa se caracteriza principalmente por la pérdida de fibras del

nervio óptico y que viene dada por el aumento de la presión intraocular (PIO) y/o la pérdida de flujo sanguíneo hacia el nervio óptico. Sin embargo, la medición de la PIO no es específica ni sensible

© El(los) autor(es) 2019. Este artículo se distribuye bajo los términos de la licencia Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), que permite su uso, distribución y reproducción sin restricciones en cualquier medio, siempre que se dé el crédito correspondiente al autor o autores originales y a la fuente, se proporcione un enlace a la licencia Creative Commons y se indique si se han realizado cambios. La renuncia a la Dedicación de Dominio Público de Creative Commons (http://creativecommons.org/publi cdomain/zero/1.0/) se aplica a los datos puestos a disposición en este artículo, a menos que se indique lo contrario.

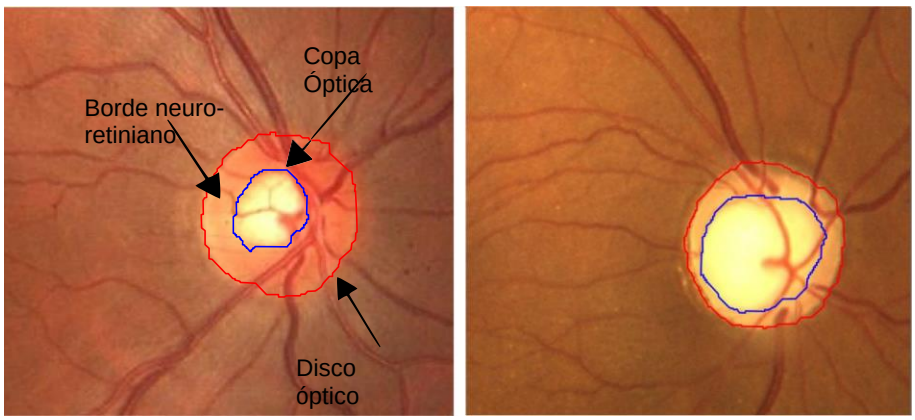

Fig. 1 Imágenes digitales de fondo de ojo recortadas alrededor del disco óptico. a Estructuras principales de un disco óptico sano y b disco óptico glaucomatoso

suficiente para ser un indicador eficaz de glaucoma, ya que puede haber daños visuales sin aumento de la PIO.

La cabeza del nervio óptico es el lugar donde los axones de las células ganglionares salen del ojo formando el disco óptico. En una imagen de fondo de ojo, el disco óptico puede separarse visualmente en dos zonas, una zona brillante y central llamada copa óptica y una parte periférica llamada borde neurorretiniano [3]. Véase la Fig. 1a.

Aunque el disco óptico (DO) y la copa están presentes en todos los individuos, un tamaño anormal de la copa con respecto al disco óptico es una característica de un ojo glaucomatoso, como se muestra en la Fig. 1b. Por esta razón, se han desarrollado diferentes enfoques para la segmentación de la copa y el disco óptico para la detección del glaucoma en imágenes de fondo de ojo en color. Algunos trabajos en la literatura se centran únicamente en la segmentación de la copa óptica y/o del disco óptico [4, 5] y otros se centran en el cálculo de la relación Copa/Disco (CDR). Esta medida se utiliza habitualmente como indicador de glaucoma, que expresa la proporción del diámetro vertical del disco óptico y la copa. Sin embargo, la medición del CDR implica un gran esfuerzo para obtener una segmentación adecuada del disco óptico y de la copa óptica.

En este artículo se presentan cinco arquitecturas de CNN diferentes para la evaluación del glaucoma. A diferencia de la mayoría de las técnicas de detección establecidas, este enfoque no necesita ninguna selección de características ni mediciones precisas de las estructuras geométricas de la cabeza del nervio óptico, como la CDR.

Antecedentes

La enfermedad del glaucoma se caracteriza principalmente por la pérdida de las fibras del nervio óptico y de los astrocitos. Esta pérdida puede examinarse midiendo el grosor del borde neurorretiniano y el tamaño de la copa óptica con respecto al disco óptico. En general, la evaluación cualitativa de la cabeza del nervio óptico, cuando se utilizan imágenes del fondo de ojo, ha sido el principal objetivo de varios trabajos en la literatura.

Por ejemplo, Wong et al. [6] presentaron un método para calcular el CDR después de obtener las máscaras de la copa óptica y del disco óptico utilizando técnicas de

conjunto de niveles. Probaron su método

en 104 imágenes y descubrieron que su método producía resultados con una variación de hasta 0,2 unidades CDR con respecto a la verdad del terreno.

Un método propuesto por Joshi et al. [7] se basa en evidencias anatómicas como las curvas de los vasos en el límite del vaso para segmentar el vaso óptico. Localizaron la copa óptica utilizando la geometría de los vasos y la transformada circular de Hough, obteniendo un error CDR de 0,12 $\pm$ 0,10. En el estudio realizado por Yin et al. [8], también utilizaron la transformada circular de Hough basada en el conocimiento para la segmentación del disco óptico y la copa óptica. Su método fue probado en 325 imágenes obteniendo un coeficiente de Dice medio de 0,92 y 0,81, respectivamente.

Otro enfoque para la segmentación del disco óptico y la copa óptica es el presentado por Cheng et al. [9], que desarrollaron una técnica para medir la CDR basada en la clasificación de superpíxeles. Evaluaron su método en 650 imágenes logrando áreas bajo la curva de 0,800 y 0,822 en dos bases de datos.

En el trabajo realizado por Díaz-Pinto et al. [10], los autores presentaron un algoritmo automático para segmentar la copa óptica y luego obtener características artesanales como el CDR, la relación copa/disco (ACDR) y la regla inferior-superior-nasal-temporal (ISNT) que comprueba el grosor del borde del disco a partir de las imágenes del fondo de ojo. Evaluaron su método en 53 imágenes obteniendo una especificidad y sensibilidad de 0,81 y 0,87 utilizando el espacio de color Luv para la segmentación del disco óptico y la copa óptica.

Un trabajo que utiliza otra información como los datos personales del paciente y la información del genoma del paciente es el presentado por Liu et al. [11]. Combinaron esa información con las imágenes del fondo de ojo obteniendo un área bajo la curva (AUC) de 0,866 para el cribado del glaucoma, que es mejor que la AUC obtenida al utilizar los datos personales individuales.

Una de las principales limitaciones de los métodos basados en características manuales (CDR, relación área copa/disco (ACDR), pliegues de vasos y regla ISNT) es la importante discrepancia en su estimación, incluso entre clasificadores humanos expertos. Por esta razón, se han centrado nuevos algoritmos en la extracción automática de características, como los métodos basados en datos [3] y las redes neuronales convolucionales (CNN).

En el artículo publicado por Bock et al. [3], proponen un método basado en datos. Este método no se basa en mediciones precisas de las estructuras geométricas de la cabeza del nervio óptico, como el CDR. En su lugar, utilizaron la idea de "Eigenimages" para extraer características que posteriormente son clasificadas por una máquina de vectores de apoyo (SVM). Evaluaron su algoritmo en 575 imágenes seleccionadas al azar del Registro de Glaucoma de Erlangen (EGR), obteniendo un AUC competitivo de 0,88. Sin embargo, las imágenes utilizadas en su trabajo son privadas y su método no puede compararse con el presentado en este trabajo.

Las redes neuronales convolucionales (CNN) fueron introducidas por primera vez por Yann LeCun [12] y son variantes inspiradas en la biología de los perceptrones multicapa. Desde entonces, se han utilizado en visión por ordenador e inteligencia artificial. Sin embargo, su relevancia no se había descubierto hasta el concurso ImageNet de 2012, en el que el objetivo principal es estimar el contenido de las imágenes naturales con el fin de realizar anotaciones automáticas utilizando un subconjunto del conjunto de datos ImageNet [13]. Su éxito se debe al uso de GPUs, rectificadores como ReLU, técnicas de aumento de datos y nuevas técnicas de regularización como Dropout [14]. El principal poder de las arquitecturas CNN reside

en su capacidad para extraer características altamente discriminantes en múltiples niveles de abstracción [15].

Las primeras capas de una CNN extraen los bordes en orientaciones y ubicaciones concretas de la imagen. Las capas intermedias detectan estructuras compuestas por arreglos particulares de

Los bordes y las últimas capas detectan estructuras más complejas que corresponden a partes de objetos familiares, o a objetos que son combinaciones de estas partes.

Entrenar una CNN desde cero no es una tarea fácil. Requieren una enorme cantidad de datos etiquetados -requisito difícil de cumplir en la tarea de evaluación del glaucoma- y recursos computacionales.

Sin embargo, existen dos alternativas para entrenar una CNN desde cero que se han aplicado previamente a varias tareas de clasificación de imágenes médicas. La primera alternativa consiste en afinar una CNN que ha sido entrenada utilizando un gran conjunto de datos etiquetados de una aplicación diferente (por ejemplo, ImageNet). Un ejemplo de esta alternativa es el trabajo de Carneiro et al. [16], donde demostraron que los modelos de CNN preentrenados en imágenes naturales, como la ImageNet, son útiles en aplicaciones de imágenes médicas, a pesar de las importantes diferencias en la apariencia de las imágenes. El estudio realizado por Chen et al. [17] demostró que el uso de una CNN preentrenada y ajustada para localizar planos estándar en imágenes de ultrasonido superó el estado del arte para el plano estándar abdominal fetal (FASP). Otro ejemplo es el estudio realizado por Tajbakhsh et al. [18], en el que llevaron a cabo un conjunto de experimentos para cuatro aplicaciones de imágenes médicas que demostraron que el uso de una CNN preentrenada funcionaba tan bien como una CNN entrenada desde cero.

La segunda alternativa consiste en utilizar una CNN entrenada en ImageNet como extractor de características, donde la CNN se aplica a una imagen de entrada y luego se extraen las características de una determinada capa oculta de la red. A continuación, las características extraídas se utilizan para entrenar un nuevo clasificador, como las máquinas de vectores de soporte (SVM), los árboles de decisión, el clasificador K-nearest-neighbor o Naive Bayes. Por ejemplo, Bar et al. [19] preentrenaron CNNs que se utilizaron como extractor de características para la identificación de patologías torácicas. Otro estudio realizado por Razavian et al. [20] demostró que utilizando las características extraídas de la red OverFeat y alimentando un clasificador SVM, es posible obtener resultados superiores en comparación con los sistemas de última generación altamente ajustados.

Para la evaluación del glaucoma, también hay varios trabajos en la literatura que emplean CNNs. Por ejemplo, Chen et al. [21] propusieron y entrenaron desde cero una arquitectura de CNN que contiene seis capas: cuatro capas convolucionales y dos capas totalmente conectadas; para clasificar automáticamente imágenes de fondo de ojo glaucomatosas. Realizaron los experimentos en dos bases de datos privadas: ORIGA-(light) que contiene 650 imágenes y SCES que contiene 1676 imágenes, logrando un AUC de 0,831 y 0,887 respectivamente. Para la base de datos ORIGA, entrenaron su arquitectura CNN seleccionando aleatoriamente 99 imágenes y utilizando las 551 restantes para la prueba. Para la base de datos SCES, utilizaron las 650 imágenes de la base de datos ORIGA para el entrenamiento, y las 1676 imágenes de la base de datos SCES para la prueba. La principal desventaja es el desequilibrio de los datos. La base de datos ORIGA está compuesta por 168 imágenes de fondo de ojo glaucomatosas y 482 normales, y la base de datos SCES contiene 1676 imágenes de fondo de ojo, de las cuales sólo 46 son glaucomatosas. Otra limitación de este trabajo es que los resultados obtenidos son difíciles de reproducir porque las bases de datos ORIGA y SCES no están disponibles públicamente.

Un estudio realizado por Alghamdi et al. [22] utilizó ocho bases de datos (cuatro públicas y cuatro privadas) para detectar anomalías en el disco óptico. Desarrollaron

un nuevo enfoque utilizando dos CNN: una CNN fue entrenada para clasificar primero la región del disco óptico y la otra CNN para clasificar la región del disco óptico en clases normal, sospechosa y anormal. Sin embargo, las cuatro bases de datos públicas (DRIVE, STARE, DIARETDB1 y

MESSIDOR) utilizadas en el trabajo de Alghamdi et al. no pueden utilizarse para la clasificación del glaucoma porque se tomaron con fines diferentes. Esto significa que esas imágenes no tienen ningún signo de glaucoma o no tienen anotaciones de glaucoma. Las bases de datos con anotaciones de glaucoma que utilizaron son privadas y, por esa razón, es difícil reproducir los resultados presentados en su trabajo.

En el estudio realizado por Abbas [23], desarrolló e implementó un sistema conocido como Glaucoma-Deep. Este sistema consiste en una arquitectura CNN no supervisada que extrae automáticamente características de las imágenes del fondo de ojo. Posteriormente, utiliza un modelo de red de creencia profunda (DBN) para seleccionar las características más discriminatorias. En su trabajo, Qaisar Abbas utiliza cuatro bases de datos para probar su método, tres de ellas públicas y una privada. Aunque su trabajo muestra buenos resultados (especificidad: 0,9801 y sensibilidad: 0,8450), no se dan detalles de la CNN ni de su arquitectura.

Merece la pena mencionar el trabajo realizado por Orlando et al. [24], en el que mostraron cómo dos CNN diferentes, OverFeat y VGG-S, podían utilizarse como extractores de características. También investigaron cómo se comportaba el rendimiento de estas redes cuando se aplicaba la ecualización adaptativa del histograma limitada por el contraste (CLAHE) y la eliminación de vasos a las imágenes del fondo de ojo. En su trabajo, utilizaron la base de datos Drishti-GS1 para probar el rendimiento de las CNN ajustadas. Observaron que la CNN OverFeat se comportó mejor que la VGG-S, obteniendo un AUC de 0,7626 y 0,7180, respectivamente. La principal limitación de este trabajo es el reducido número de imágenes (101 imágenes) utilizadas para probar el rendimiento de las CNN. Sin embargo, su método logró una puntuación AUC competitiva con respecto a otras estrategias existentes.

En este trabajo se presenta un análisis de cinco arquitecturas CNN diferentes entrenadas en ImageNet y utilizadas como clasificadores de glaucoma. Se pusieron a punto y se probaron utilizando exclusivamente bases de datos públicas, lo que difiere de la mayoría de los trabajos presentados en la literatura que utilizan bases de datos privadas. La alta precisión, especificidad y sensibilidad obtenida de este análisis sugiere que las arquitecturas CNN entrenadas en ImageNet son una alternativa robusta para un algoritmo de detección automática de glaucoma. Estas CNNs funcionan correctamente en imágenes de fondo de ojo en color pertenecientes a cinco bases de datos públicas diferentes (1707 imágenes) con alto grado de variabilidad. Además, introducimos una nueva base de datos pública, ACRIMA, compuesta por 705 imágenes etiquetadas (396 glaucomatosas y 309 normales), que podría utilizarse como banco de pruebas para realizar más comparaciones entre los métodos desarrollados para la clasificación del glaucoma.

Material y métodos

ACRIMA: una nueva base de datos pública

Hay pocas bases de datos disponibles públicamente con imágenes etiquetadas para el glaucoma que puedan utilizarse para la evaluación de los métodos de clasificación del glaucoma. Por ese motivo, los autores se complacen en presentar una nueva base de datos con etiquetado de glaucoma disponible al público, denominada ACRIMA. [1] Las imágenes de esta base de datos proceden del proyecto ACRIMA (TIN2013- 46751-R) fundado por el Ministerio de Economía y Competitividad de España, cuyo objetivo es el desarrollo de algoritmos automáticos para la evaluación de enfermedades de la retina.

[1] Enlace a la base de datos ACRIMA. Estará disponible públicamente después de que se acepte el documento.

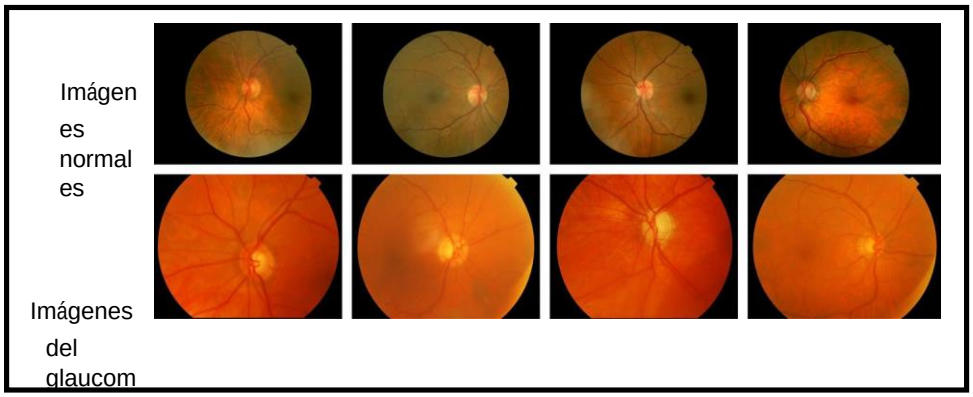

Fig. 2 Ejemplos de la nueva base de datos de acceso público. Imágenes de fondo de ojo normales y de glaucoma de la nueva base de datos disponible al público (ACRIMA)

Tabla 1 Lista de todas las bases de datos públicas disponibles con etiquetas de glaucoma

Base de datos	Glaucoma	Normal	Total
HRF [25]	27	18	45
Drishti-GS1 [4]	70	31	101
RIM-ONE [26]	194	261	455
sjchoi86-HRF [23, 27]	101	300	401
ACRIMA	*396*	*309*	*705*
	788	919	1707

La cursiva representa la nueva base de datos de acceso público

Descripción de la base de datos

La base de datos ACRIMA está compuesta por 705 imágenes de fondo de ojo (396 glaucomatosas y 309 nor- malas). Forman parte del proyecto ACRIMA y se obtuvieron de pacientes glaucomatosos y normales con su consentimiento previo y de acuerdo con las normas éticas establecidas en la Declaración de Helsinki de 1964. Todos los pacientes fueron seleccionados por expertos basándose en sus criterios y en los hallazgos clínicos durante el examen. La mayoría de las imágenes de fondo de ojo de esta base de datos se tomaron del ojo izquierdo y derecho previamente dilatado y centrado en el disco óptico. Algunas de ellas se descartaron debido a los artefactos, el ruido y el escaso contraste. Se capturaron utilizando la cámara de retina Topcon TRC y el sistema de captura IMAGEnet®. Las imágenes se tomaron con un campo de visión de 35°.

Todas las imágenes de la base de datos ACRIMA fueron anotadas por dos expertos en glaucoma con 8 años de experiencia. No se tuvo en cuenta ninguna otra información clínica a la hora de etiquetar las imágenes. Esta primera versión de la base de datos ACRIMA sólo podía utilizarse para tareas de clasificación. No se proporciona la segmentación del disco óptico ni de la copa óptica. En la Fig. 2 se muestran ejemplos de imágenes de la base de datos ACRIMA.

Otras bases de datos

Además de la base de datos ACRIMA, en este trabajo se utilizaron otras cuatro bases

de datos públicas: La base de datos HRF [25], que contiene 45 imágenes; la base de datos Drishti-GS1 [4], que consta de 101 imágenes; la base de datos RIM-ONE [26], que está compuesta por 455 imágenes; y la base de datos sjchoi86- HRF [27], que está compuesta por 401 imágenes. Todas estas bases de datos se muestran en detalle en la Tabla 1.

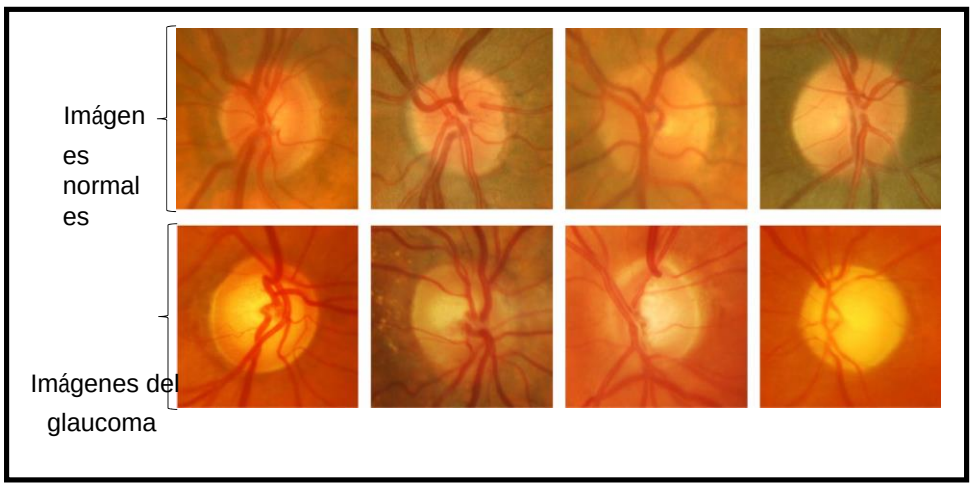

Fig. 3 Ejemplos de imágenes de fondo de ojo recortadas. Imágenes de fondo de ojo recortadas utilizadas para afinar y probar las CNN

Para todos los experimentos realizados en este trabajo, se utilizó la biblioteca de aprendizaje profundo de código abierto Keras [28] y la GPU NVIDIA Titan V. La librería Keras es una forma sencilla de utilizar, implementar y ajustar arquitecturas CNNs construidas sobre Theano, TensorFlow o CNTK.

Preprocesamiento

Las imágenes de fondo de ojo utilizadas para el proceso de ajuste se recortaron automáticamente alrededor del disco óptico utilizando un cuadro de delimitación de 1,5 veces el radio del disco óptico. Excepto la base de datos RIM-ONE, que se recortó originalmente alrededor del disco óptico. Para realizar este recorte, empleamos el método propuesto en [29]. En su método, Xu et al. utilizaron una CNN básica para encontrar los píxeles más probables en la región del disco óptico. A continuación, clasifican esos píxeles candidatos mediante el uso de un umbral.

El recorte de las imágenes alrededor del disco óptico tiene una razón clínica, ya que la enfermedad del glaucoma afecta principalmente al disco óptico y sus alrededores. Además, Orlando et al. [24] demostraron que el recorte de las imágenes alrededor del disco óptico resultaba más eficaz que el uso de las imágenes completas cuando se utilizaba la CNN para la evaluación del glaucoma. En la Fig. 3 se muestran ejemplos de las imágenes utilizadas para el ajuste de las CNN.

Arquitecturas CNN entrenadas en ImageNet

En este trabajo, las arquitecturas VGG16 [15], VGG19 [15], InceptionV3 [30, 31], ResNet50 [32] y Xception [33] se ajustaron a la tarea de evaluación del glaucoma utilizando sus versiones entrenadas en ImageNet disponibles en el núcleo de Keras. Para utilizar estas redes en esta tarea, se cambió la última capa totalmente conectada de cada CNN por una capa de pooling global (GlobalAveragePooling2D) seguida de una capa totalmente conectada de dos nodos que representaban dos clases (glaucoma y sano) y un clasificador softmax. Por lo tanto, contando las nuevas capas superiores en cada CNN, el número total de capas Keras en las arquitecturas de red VGG16 y VGG19 era de 20 y 23, respectivamente. La arquitectura InceptionV3 está compuesta por 312 capas Keras y las arquitecturas ResNet50 y Xception están compuestas por 176

y 133 capas Keras, respectivamente. Obsérvese que, para afinar los modelos, las imágenes se recortaron automáticamente alrededor del disco óptico, como se ha mencionado anteriormente.

Para obtener el mejor rendimiento de cada modelo, realizamos varios experimentos variando el número de capas ajustadas y el número de épocas. En primer lugar, para el número de capas ajustadas, comenzamos ajustando la última capa ponderada de las arquitecturas CNN, manteniendo las demás capas en modo "no entrenable". Después, se aumentó el número de capas afinadas hasta actualizar todas las capas de la CNN.

El segundo experimento consistió en analizar el impacto del número de épocas que presentan el mejor rendimiento de cada arquitectura. Otros hiperparámetros, como el tamaño del lote, la tasa de aprendizaje, etc., se fijaron mientras se variaba el número de capas afinadas y el número de épocas. Por ejemplo, el número de épocas se fijó en 100, para el primer experimento, y el número de capas en modo "no entrenable" se fijó en 0 para el segundo experimento. Para ambos experimentos, se utilizó el Descenso Gradiente Estocástico (SGD) como optimizador, el tamaño del lote se fijó en 8, la tasa de aprendizaje en 1e-4 y el impulso en 0,9. Todos estos hiperparámetros se eligieron de forma óptima para obtener el mejor rendimiento en nuestro conjunto de experimentos.

Además de estos experimentos, también se evaluó el rendimiento de las CNNs (VGG16, VGG19, Inceptionv3, ResNet50 y Xception) utilizando la técnica de k-fold Cross Validation con $k = 10$, siguiendo el procedimiento descrito en [34]. Para evitar el sobreajuste y aumentar la robustez de los modelos, se aumentaron las imágenes disponibles utilizando rotaciones aleatorias, zoom en un rango entre 0 y 0,2 y volteo horizontal y vertical. Las imágenes también se redimensionaron al tamaño de entrada por defecto de cada arquitectura CNN (224×224 para VGG16, VGG19 y ResNet50 y 299×299 para Incep- tionv3 y Xception).

También se llevó a cabo una evaluación particular del rendimiento de las CNN, utilizando conjuntos de datos que no se utilizaron durante la etapa de entrenamiento. A diferencia de la mayoría de los trabajos en la literatura, este experimento comprueba el rendimiento de las CNNs en bases de datos completas que el sistema no ha visto durante la etapa de entrenamiento.

El experimento final es la comparación entre la mejor de las cinco CNN mencionadas anteriormente con un algoritmo de última generación que también utiliza bases de datos públicas.

Resultados experimentales y discusión

Como se mencionó en la sección anterior, se llevaron a cabo dos experimentos iniciales para ver el efecto de la sintonización fina de diferentes números de capas y el número de épocas. En la Fig. 4 se puede ver la tendencia de mejora desde la sintonización superficial hasta la sintonización profunda para la tarea de evaluación del glaucoma de cada CNN. El eje x representa el número de capas en modo "no entrenable" de cada CNN. El experimento comienza con el ajuste fino de todas las capas (eje $x = 0$) hasta el ajuste fino de la última capa entrenable del modelo.

A partir de este experimento, podemos ver que el ajuste fino de todas las capas entrenables de la CNN, o hacer un ajuste profundo, es la mejor opción cuando se trata de obtener el mejor rendimiento, como también demostraron Tajbakhsh et al. [18] en su artículo.

Para el otro experimento inicial, ponemos en modo entrenable todas las capas y vemos el rendimiento de la CNN cuando se ajusta con precisión de 1 a 250 épocas. Es importante destacar que la evaluación del rendimiento de este experimento se realizó sobre el conjunto de validación. El resultado de este experimento puede verse en la Fig.

5. A partir de este experimento, podemos ver que alrededor de 200 es el número óptimo de épocas para obtener el mejor rendimiento para el proceso de ajuste fino en nuestro conjunto de experimentos.

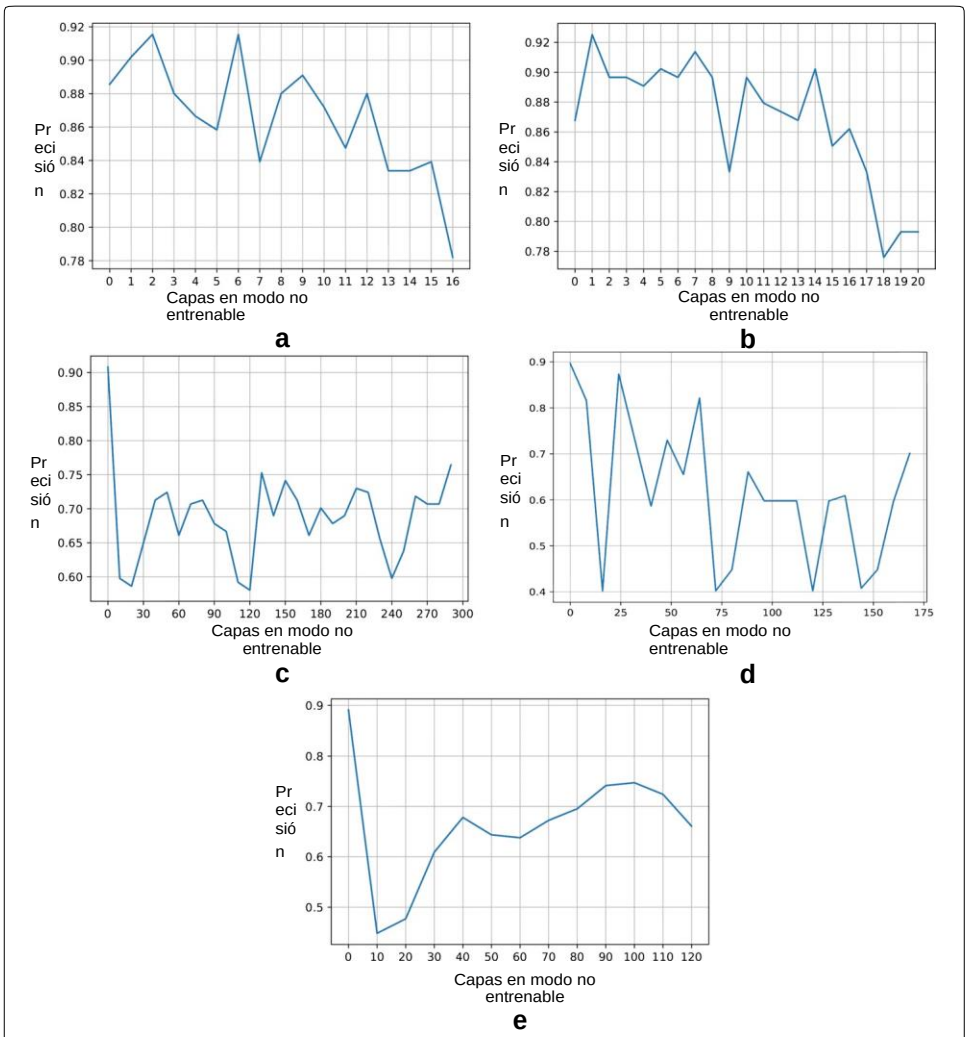

Fig. 4 Gráfico que muestra el efecto del ajuste fino de un número diferente de capas. a VGG16, b VGG19, c InceptionV3, d ResNet50 y la arquitectura e Xception

Para la evaluación del modelo, se realizó una validación cruzada de 10 veces. Por tanto, se obtuvieron 10 valores de área bajo la curva (AUC), precisión, especificidad, sensibilidad y puntuación F. A continuación, se calculó la media y la desviación estándar de estos valores para cada arquitectura CNN. Los resultados de cada modelo ajustado se presentan en la Tabla 2. Además del área bajo la curva (AUC), la precisión, la especificidad, la sensibilidad y la puntuación F, también calculamos para cada pliegue el valor P mediante la prueba U de Mann-Whit- ney [35, 36] y utilizamos la técnica de bootstrapping para calcular el intervalo de confianza de los valores AUC. El valor P se calculó comparando la distribución de la verdad básica y la distribución de las etiquetas obtenidas de cada modelo. La distribución de etiquetas de cada modelo se obtuvo mediante el umbral de las probabilidades obtenidas de cada CNN (glaucoma si la puntuación es > 0,5, normal en caso contrario). La idea de proporcionar intervalos de confianza para cada CNN es proporcionar una probabilidad de que el AUC de cada modelo caiga dentro del rango cuando se hagan predicciones sobre nuevos datos. Para calcular estos intervalos, utilizamos los modelos entrenados en cada pliegue para predecir imágenes seleccionadas al azar

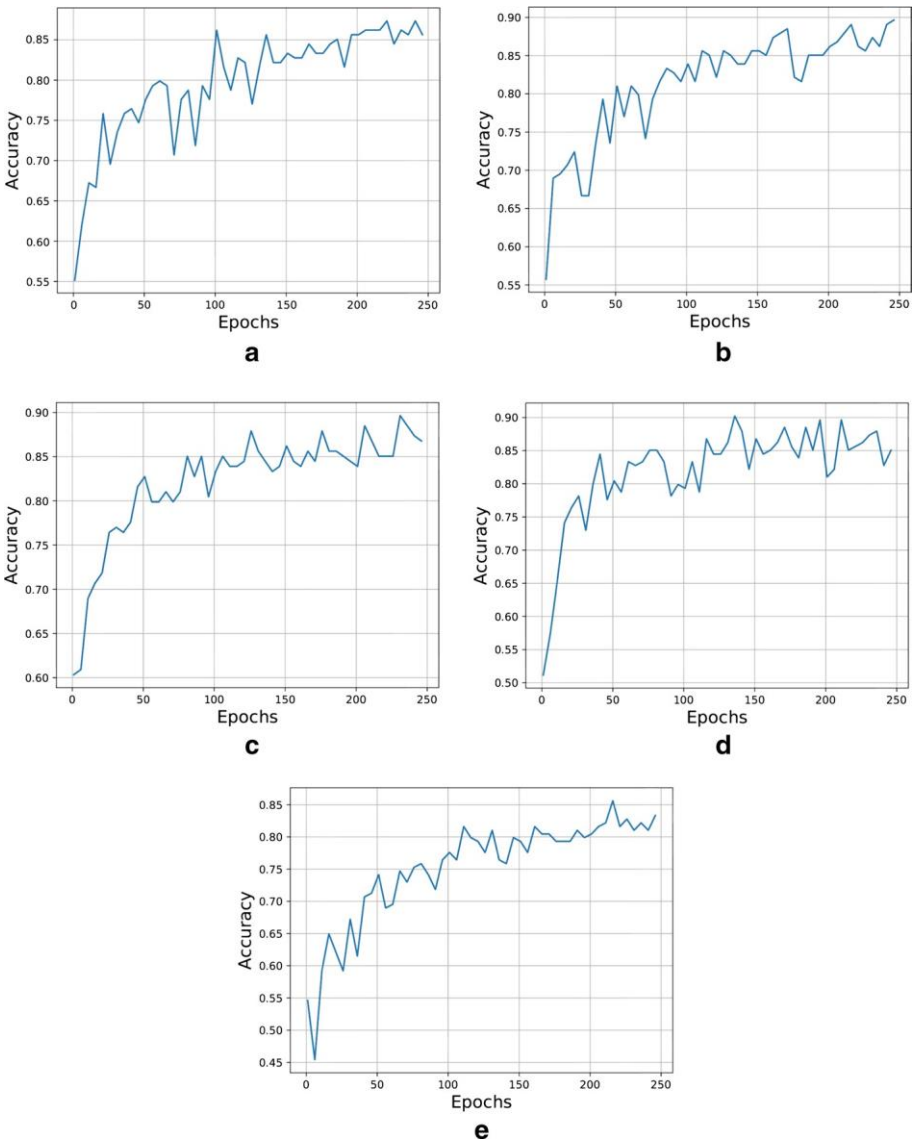

Fig. 5 Gráfico que muestra el efecto del número de épocas. a VGG16, b VGG19, c InceptionV3, d ResNet50 y e Arquitectura de Xception

del conjunto de pruebas. Se realizaron diez repeticiones bootstrap para cada pliegue, es decir, 100 repeticiones para cada modelo.

Además de las mediciones indicadas en la Tabla 2, en la Fig. 6 se han representado las curvas ROC para la especificidad y la sensibilidad medias, obtenidas mediante una validación cruzada de 10 veces. Se puede observar que todos los modelos propuestos con ajuste fino tienen un rendimiento realmente bueno para la tarea de evaluación del glaucoma. Como el rendimiento de los modelos es comparable, las características de la CNN, como el número de parámetros, pueden utilizarse para determinar qué modelo es mejor que los demás.

Para comparar los resultados obtenidos con otros trabajos de la literatura, implementamos, entrenamos y probamos sobre las mismas imágenes las redes neuronales propuestas

Tabla 2 Resultados de cada modelo realizando un ajuste profundo y una validación cruzada de 10 veces

Nombre del modelo	AUC	Intervalo de confianza del 95%	Precisión	Especificidad	Sensibilidad	Fscore	Valor P
VGG16 [15]	0.9632 (0.0149)	95.81-96.87%	0.8948 (0.0253)	0.8816 (0.0612)	0.9057 (0.0331)	0.9005 (0.0231)	0.3240
VGG19 [15]	0.9686 (0.0158)	96.45-97.39%	0.9069 (0.0318)	0.8846 (0.0362)	0.9240 (0.0434)	0.9125 (0.0312)	0.3607
InceptionV3 [30]	0.9653 (0.0135)	96.12-97.32%	0.9000 (0.0201)	0.8752 (0.0358)	0.9216 (0.0311)	0.9056 (0.0236)	0.3126
ResNet50 [32]	0.9614 (0.0171)	95.62-96.77%	0.9029 (0.0249)	0.8943 (0.0350)	0.9105 (0.0282)	0.9076 (0.0251)	0.3885
Xcepción [33]	0.9605 (0.0170)	95.92-97.07%	0.8977 (0.0264)	0.8580 (0.0398)	0.9346 (0.0247)	0.9051 (0.0274)	0.2729

por Chen et al. [21] y Alghamdi et al. [22]. Los resultados obtenidos con estos modeles se presentan en la Fig. 6. Aunque obtuvieron una elevada área bajo la curva ROC con sus métodos, los sistemas propuestos en este trabajo los superan claramente.

En la Tabla 3, se muestra el número de parámetros y el AUC obtenido de cada arquitectura CNN. Se puede observar que, aunque VGG16 y VGG19 presentan un AUC más alto que la arquitectura Xception, tienen muchos más parámetros que ajustar, lo que requiere más potencia de cálculo y recursos. Por lo tanto, la arquitectura Xception presenta un mejor equilibrio entre el número de parámetros y el AUC obtenido que las otras arquitecturas.

En las Figs. 7 y 8 se presentan muestras de los resultados de clasificación obtenidos con la arquitectura Xception, incluyendo ejemplos de clasificación correcta e incorrecta. Los valores de puntuación bajos (cercanos a 0) significan una clasificación errónea y los valores de puntuación cercanos a 1 significan una clasificación correcta.

Creemos que una posible razón por la que la CNN clasifica erróneamente las imágenes glaucomatosas es porque no tienen la gran zona brillante dentro del disco óptico. Como se ha mencionado anteriormente, la pérdida del nervio óptico es, la mayoría de las veces, visible desde la imagen del fondo de ojo. La zona más brillante (o copa óptica) dentro del disco óptico suele ser mayor en una imagen glaucomatosa. Parece que la CNN aprendió a reconocer las imágenes glaucomatosas basándose en esta

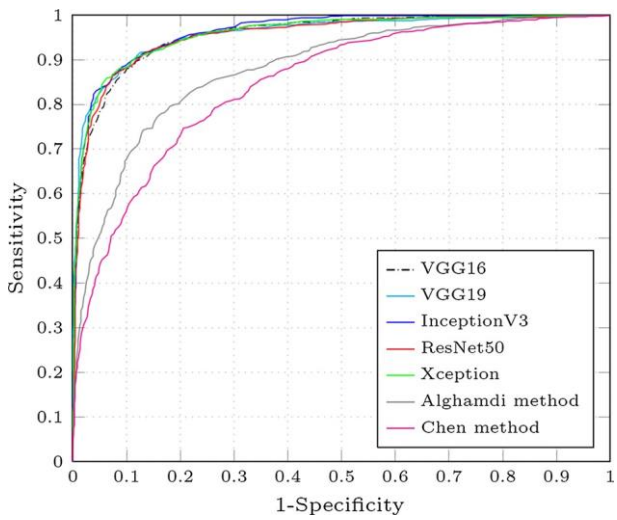

Fig. 6 Curvas ROC medias de cada CNN. Curvas ROC medias para cada arquitectura CNN ajustada (VGG16 [15], VGG19 [15], InceptionV3 [31], ResNet50 [32] y Xception [33]), el método de Alghamdi [22] y Chen [21]

Tabla 3 Número de parámetros y AUC obtenidos para cada arquitectura

Nombre del modelo	# Parámetros (en millones)	AUC
VGG16	138	0.9632 (0.0149)
VGG19	144	0.9686 (0.0158)
InceptionV3	23	0.9653 (0.0135)
ResNet50	25	0.9614 (0.0171)
Xception	*22*	*0.9605 (0.0170)*

La mejor arquitectura en términos de AUC y número de parámetros (en cursiva)

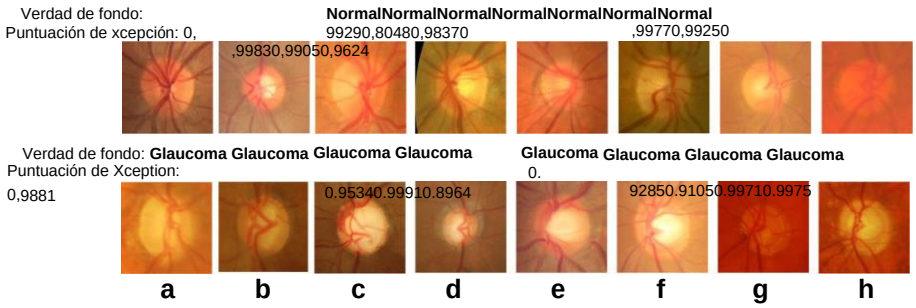

Fig. 7 Ejemplos de clasificación correcta de imágenes normales y de glaucoma. Muestras de imágenes normales y de glaucoma bien clasificadas por la arquitectura Xception.

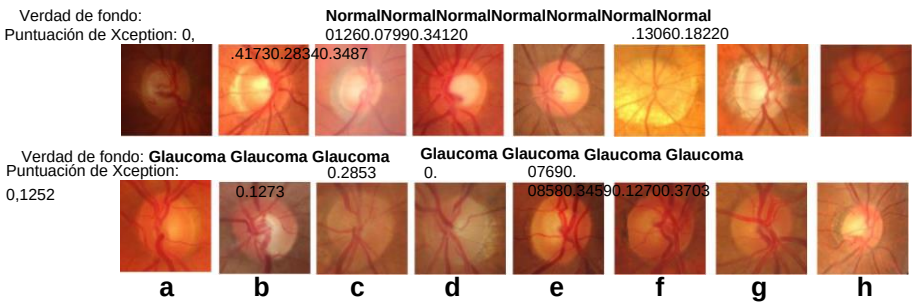

Fig. 8 Ejemplos de imágenes normales y de glaucoma mal clasificadas. Muestras de imágenes normales y de glaucoma mal clasificadas por la arquitectura Xception

criterios. Las imágenes sin una gran zona brillante se clasifican como normales (véanse las imágenes de la fila inferior de la Fig. 8). Otros posibles factores para la clasificación errónea son la baja calidad de las imágenes (véase la fila superior de la Fig. 8a, h).

Dados los cambios de iluminación, la evaluación del glaucoma mediante imágenes de fondo de ojo no es una tarea fácil. Un método desarrollado que clasifica correctamente las imágenes de una base de datos determinada no tiene por qué funcionar bien cuando se aplica a imágenes de una base de datos diferente. Un experimento crítico que evalúa el rendimiento de un clasificador de glaucoma consiste en utilizar imágenes que provienen de un sensor o base de datos diferente. Por ello, se realizaron cinco experimentos diferentes utilizando la arquitectura Xception y todas las bases de datos públicas etiquetadas de glaucoma (HRF, Drishti-GS1, RIM, sjchoi86-HRF y ACRIMA). En primer lugar, la arquitectura Xception se puso a punto utilizando todas las bases de datos excepto las imágenes que pertenecen a la base de datos que se va a probar. En segundo lugar, se prueba el modelo entrenado en la base de datos deseada. Este enfoque se repite para probar las bases de datos HRF, Drishti-GS1, RIM, sjchoi86-HRF y ACRIMA. Los resultados obtenidos en estos experimentos se presentan en la Fig. 9 y en la Tabla 4, en los que se puede ver que, aunque la arquitectura Xception se puso a punto sin utilizar imágenes de estas bases de datos, su rendimiento es prometedor.

Los intervalos de confianza de las AUC presentados en la Tabla 4 se calcularon a partir de 1.000 réplicas de la correa de arranque. Utilizamos el modelo entrenado de cada experimento y seleccionamos al azar

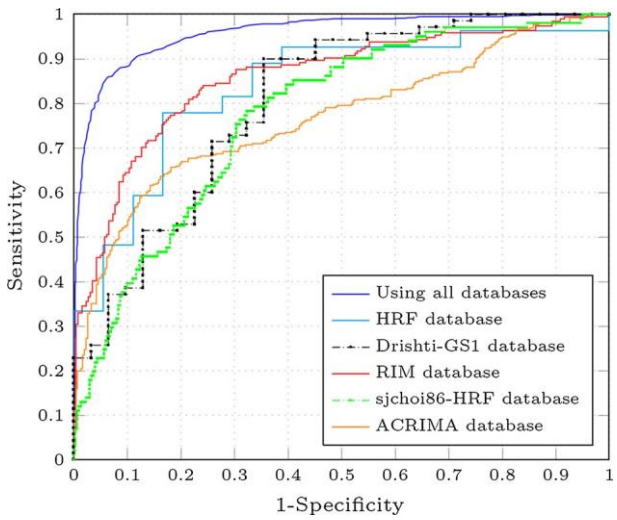

Fig. 9 Curvas ROC para la arquitectura Xception en diferentes experimentos. Curvas ROC para la arquitectura Xception en diferentes experimentos. Cuando se utilizan las bases de datos HRF, Drishti-GS1, RIM, sjchoi86-HRF y ACRIMA únicamente como conjunto de pruebas y cuando se utilizan todas las bases de datos

Tabla 4 Resultados obtenidos para HRF, Drishti-GS1, RIM, sjchoi86-HRF y la nueva base de datos pública ACRIMA, utilizando la arquitectura Xception representada en AUC, intervalo de confianza de AUC, precisión, sensibilidad y especificidad

Base de datos	AUC	AUC's 95% intervalo de confianza	Precisión	Sensibilidad	Especificidad	# imágenes
HRF	0.8354	50.00-100.00%	0.8000	0.8333	0.7778	45
Drishti-GS1	0.8041	50.49-92.55%	0.7525	0.7419	0.7143	101
RIM-ONE	0.8575	77.53-91.12%	0.7121	0.7931	0.7990	455
sjchoi86-HRF	0.7739	66.44-86.85%	0.7082	0.7033	0.7030	401
ACRIMA	0.7678	68.41-81.81%	0.7021	0.6893	0.7020	705

imágenes del conjunto de pruebas para obtener un valor AUC para cada réplica bootstrap. En la Fig. 10 se puede ver la distribución de los 1000 valores AUC.

En la Tabla 4 se puede ver cómo la precisión de la CNN disminuye (aproximadamente un 15%) cuando predice imágenes de bases de datos diferentes a las utilizadas para el entrenamiento. Esta caída en la precisión muestra que la CNN no generaliza bien cuando clasifica imágenes de diferentes bases de datos. Esto se debe a que las imágenes de diferentes bases de datos se etiquetan de dos formas diferentes: la primera es cuando los expertos se basan en el historial médico del paciente y en la propia imagen del fondo de ojo para asignar una etiqueta y la otra forma es cuando los expertos se basan únicamente en la información visual de la imagen del fondo de ojo. La última forma de etiquetar las imágenes aumenta el ruido en las etiquetas y dificulta aún más la generalización de un sistema de clasificación automática del glaucoma. Teniendo en cuenta que la CNN sólo se basa en la información bruta de los píxeles para clasificar las imágenes, es de esperar que la precisión se vea drásticamente afectada al probar el sistema con imágenes de diferentes bases de datos. Diferentes bases de datos significan diferentes sistemas de etiquetado.

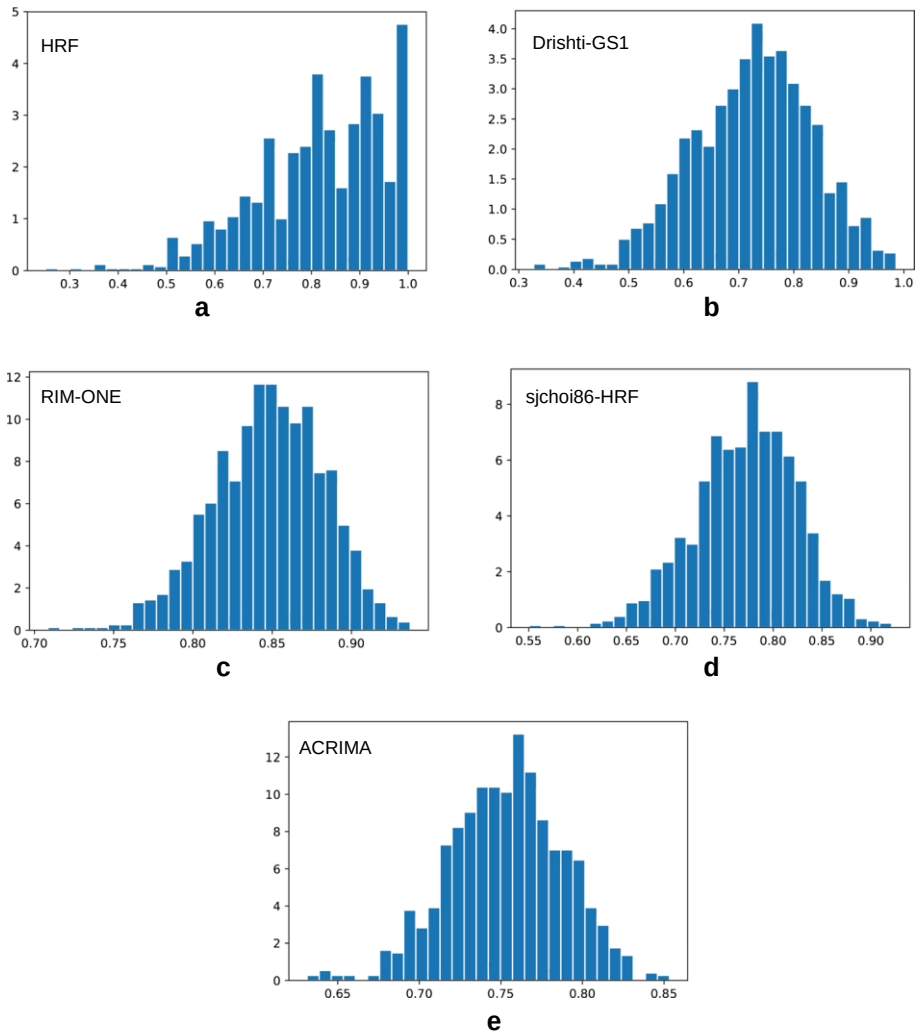

Fig. 10 Distribución de los valores AUC por base de datos. Histograma de los 1000 valores AUC al evaluar la arquitectura Xception mediante la técnica de bootstrap. Muestra los resultados de las repeticiones de bootstrap al utilizar a HRF, b Drishti-GS1, c RIM-ONE, d sjchoi86-HRF y e ACRIMA como única prueba

Analizando en profundidad los resultados de la Tabla 4, se pueden destacar dos resultados principales. El primero consiste en comprobar el rendimiento real de las CNN cuando se prueba el modelo con imágenes de bases de datos que no se utilizaron durante la fase de entrenamiento. No se trata sólo de datos no vistos, sino de bases de datos completas con diferentes etiquetadores expertos y características de las imágenes. A diferencia de la mayoría de los trabajos de la literatura, en los que se prueban sus sistemas con datos no vistos pero utilizando imágenes de la misma base de datos con una apariencia similar.

El segundo resultado es mostrar la dificultad de este problema. La escasez de imágenes etiquetadas, y el hecho de que las imágenes disponibles procedan de diferentes bases de datos, hace que el desarrollo de un clasificador de glaucoma robusto que funcione para cada base de datos sea un problema complejo.

Una posible solución para aumentar el rendimiento de las CNN es utilizar parte de los datos para el entrenamiento y parte para la prueba. De este modo, podríamos probar las CNN en datos no vistos

pero utilizando imágenes de las bases de datos utilizadas durante la etapa de entrenamiento. Otra solución podría ser el reentrenamiento de las CNN al intentar clasificar imágenes de bases de datos diferentes a las utilizadas para el entrenamiento.

Hemos comprobado que este descenso de la precisión no se produce cuando se entrena la CNN con el 70% de los datos (1195 imágenes) y se prueba con el 30% de los datos (512 imágenes) de las mismas bases de datos. De este experimento, obtuvimos un AUC de 0,9464, una precisión de 0,8908, una sensibilidad de 0,9175 y una especificidad de 0,8571. Esto demuestra que la CNN funciona bien a la hora de clasificar imágenes no vistas de las mismas bases de datos que se utilizaron para el entrenamiento.

Gracias a la disponibilidad pública de la base de datos Drishti-GS1, también es posible realizar una comparación con otros algoritmos de última generación que utilizaron esta base de datos. Por ejemplo, en el trabajo desarrollado por Chakravarty et al. [37], obtuvieron un AUC de 0,78 cuando probaron su método en esta base de datos. Otro ejemplo es el trabajo presentado por Orlando et al. [24], en el que obtuvieron un AUC de 0,76 utilizando CNNs preentrenadas aplicadas a la base de datos Drishti-GS1. En la Tabla 4 se puede observar que el método propuesto en este trabajo supera (AUC = 0,8041) a los trabajos existentes. Además, hay que tener en cuenta que Chakravarty et al. [37] y Orlando et al. [24] evaluaron sus métodos utilizando la misma base de datos para el entrenamiento y la prueba, a diferencia de los experimentos realizados en este trabajo en los que la base de datos Drishti-GS1 sólo se utiliza para la prueba. Esta podría ser la razón de la baja ganancia de nuestro método con respecto a los demás. Esta ganancia aumenta significativamente (AUC = 0,9605 para Xception. Véase la Tabla 3) cuando se utiliza la base de datos Drishti-GS1 para ambas etapas, entrenamiento y prueba.

Se realizó un experimento adicional utilizando la arquitectura Xception y sólo la base de datos Drishti-GS1. Para hacer una comparación justa con el método publicado en [24], las imágenes de esta base de datos se dividieron aleatoriamente en entrenamiento, validación y prueba, como se hizo en su trabajo: El 70% para el entrenamiento, el 30% para la prueba y el 30% de las imágenes de entrenamiento se seleccionaron para el conjunto de validación. Para este experimento, el tamaño del lote se redujo a 1 y la tasa de aprendizaje a 1e-3 debido al reducido número de imágenes de la base de datos Drishti-GS1 (101 imágenes en total). Con esta configuración, se obtuvo un AUC de 0,8261 con un intervalo de confianza del 95% de 59,71 a 95,20%. Además, utilizamos la técnica de bootstrapping y el modelo entrenado para obtener 1000 valores de AUC y comprobar su distribución (véase la Fig. 11).

Los experimentos anteriores, el análisis de esas CNNs y los resultados obtenidos demuestran ser un avance importante para la tarea de clasificación del glaucoma utilizando imágenes de fondo de ojo.

Tiempo de cálculo

Para este trabajo hemos utilizado la librería Keras y una GPU Titan Xp. Medimos el tiempo consumido por cada modelo con la configuración descrita anteriormente (técnica de regularización, número de capas de ajuste fino, número de imágenes para el entrenamiento, tamaño del lote, etc). El tiempo medio de ajuste es de 1 h y 40 min para el VGG16, 1 h y 55 min para el VGG19, 1 h y 45 min para el InceptionV3, 1 h y 15 min para el ResNet50 y 2 h y 40 min para la arquitectura Xception. El tiempo de cada arquitectura se obtuvo promediando el tiempo consumido por cada pliegue durante la etapa de ajuste fino. Una vez ajustados los modelos, se necesitan 46 ms para asignar

una probabilidad de glaucoma a cada imagen de la retina.

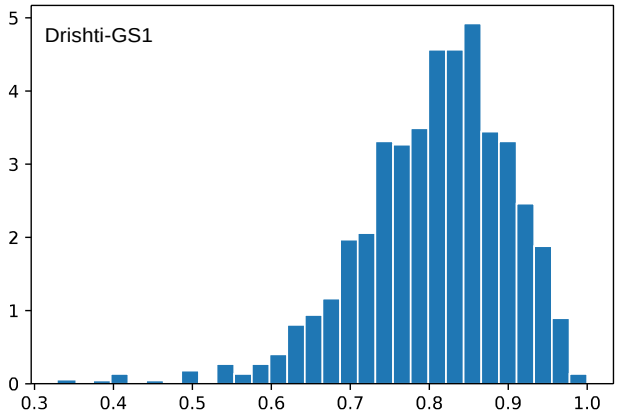

Fig. 11 Distribución de los valores AUC para la base de datos Drishti-GS1. Histograma de los 1.000 valores AUC obtenidos a partir de la técnica de bootstrapping que evalúa la arquitectura Xception entrenada y probada en la base de datos Drishti-GS1.

Conclusión

En este trabajo se analizaron cinco arquitecturas de CNN entrenadas en ImageNet (VGG16, VGG19, InceptionV3, ResNet50 y Xception) y se utilizaron como clasificadores de glaucoma. Utilizando únicamente bases de datos públicas, la arquitectura Xception muestra el mejor rendimiento para la clasificación del glaucoma, que se evaluó como el compromiso entre el AUC y el número de parámetros de la CNN. A partir de las 1.707 imágenes y de la técnica de aumento de datos, se obtuvo un AUC medio de *0,9605* con un intervalo de confianza del 95% del 95,92-97,07%, una especificidad media de *0,8580* y una sensibilidad media de *0,9346* después de afinar la arquitectura Xception, mejorando significativamente otros trabajos del estado del arte.

Además, un análisis adicional muestra que el modelo ajustado tiene un rendimiento competitivo cuando se prueba con imágenes que provienen de una base de datos completamente diferente. Este experimento difiere del enfoque habitual en el que se utiliza un subconjunto de una base de datos para el entrenamiento y el otro subconjunto para las pruebas. Utilizando sólo la base de datos ACRIMA como conjunto de pruebas, se obtuvo un AUC de 0,7678 con un intervalo de confianza del 95% de 68,41 a 81,81%. El mismo experimento se realizó para las otras cuatro bases de datos públicas: HRF, Drishti-GS1, RIM-ONE, sjchoi86-HRF, obteniendo un AUC de 0,8354, 0,8041, 0,8575 y 0,7739, respectivamente.

La base de datos ACRIMA2 está compuesta por 396 imágenes glaucomatosas y 309 imágenes normales

y podría utilizarse fácilmente como banco de pruebas para nuevas comparaciones y/o análisis. Los autores animan a la comunidad científica a probar sus modelos utilizando la nueva base de datos disponible públicamente y a comparar sus resultados con el método propuesto en este artículo.

Como trabajo adicional, el uso de imágenes sintéticas para el entrenamiento de las CNN podría ser de gran interés para aumentar el número de imágenes de entrenamiento. De este modo, podríamos entrenar clasificadores de glaucoma aún más robustos.

[2] Enlace a la base de datos ACRIMA durante el proceso de revisión.

Limitaciones del estudio

Aunque utilizamos una base de datos bastante grande y usamos el aumento de datos para afinar las CNN, sigue habiendo una limitación cuando se trata de generalizar. Como se muestra en la Tabla 4, el rendimiento disminuye cuando se prueba la CNN en bases de datos diferentes de las utilizadas para el entrenamiento. Además de este problema, encontramos que los diferentes criterios de etiquetado son otro problema al que nos enfrentamos al desarrollar sistemas de evaluación automática del glaucoma. La mayoría de las bases de datos disponibles públicamente difieren en la forma de etiquetarlas, en la información que los expertos clínicos utilizan para evaluar las imágenes y en la calidad de las imágenes del fondo de ojo.

Contribuciones de los autores

ADP redactó el artículo. SM, VN, TK, JMM y AN proporcionaron apoyo técnico y de redacción científica a este manuscrito. Todos los autores leyeron y aprobaron el manuscrito final.

Detalles del autor

[1] Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, España. [2] Pattern Recognition Lab, University of Erlangen-Nuremberg, Erlangen, Germany.[3] iTEAM, Univer- sitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain. [4] Instituto de Ciencias Biomédicas, Universidad CEU Cardenal Herrera, Avenida del Seminario s/n, Moncada, 46313 Valencia, España.

Agradecimientos

Agradecemos el apoyo de NVIDIA Corporation con la donación de la GPU Titan V utilizada para esta investigación.

Intereses contrapuestos

Los autores declaran que no tienen intereses contrapuestos.

Disponibilidad de datos y materiales

Los conjuntos de datos, el código fuente y las ponderaciones de las CNN utilizadas y/o analizadas durante el presente estudio están disponibles públicamente en este enlace https://figshare.com/s/c2d31f850af14c5b5232.

Consentimiento para la publicación

No se aplica.

Aprobación ética y consentimiento para participar

No se aplica

Financiación

Este trabajo ha sido apoyado por el Ministerio de Economía y Competitividad de España, Proyecto ACRIMA [TIN2013-46751-R] y el Proyecto GALAHAD [H2020-ICT-2016-2017, 732613]. En particular, el trabajo de Andrés Díaz-Pinto ha sido apoyado por la Generalitat Valenciana con la beca Santiago Grisolía [GRISOLIA/2015/027]. (Autor correspondiente: Andrés Díaz-Pinto).

Nota del editor

Springer Nature se mantiene neutral con respecto a las reclamaciones jurisdiccionales en los mapas publicados y las afiliaciones institucionales.

Recibido: 16 de octubre de 2018 Aceptado: 13 de marzo de 2019
Published online: 20 March 2019

Referencias

1. Organización Mundial de la Salud. Boletín de la Organización Mundial de la Salud, Vol 82(11). 2004. http://www.who.int/bulletin/ volumes/82/11/en/infocus.pdf?ua 1. Consultado el 5 de mayo de 2016.
2. Bourne RRA. El glaucoma mundial a través del espejo. Br J Ophthalmol. 2006;90:253-4. https://doi.org/10.1136/bjo.2005.083527.
3. Bock R, Meier J, Nyúl LG, Hornegger J, Michelson G. Índice de riesgo de glaucoma: detección automática de glaucoma a partir de imágenes de fondo de ojo en color . Med Image Anal. 2010;14:471–81. https://doi.org/10.1016/j.media.2009.12.006.
4. Sivaswamy J, Krishnadas SR, Joshi GD, Jain M, Ujjwal A, ST Drishti. Conjunto de datos de imágenes de la retina para la segmentación de la cabeza del nervio óptico (ONH). En: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). 2014, p. 53-6. https://doi. org/10.1109/ISBI.2014.6867807.
5. Morales S, Naranjo V, Angulo J, Alcañiz M. Detección automática de disco óptico basada en PCA y mor- fología matemática. IEEE Trans Med Imag. 2013;32:786–96. https://doi.org/10.1109/TMI.2013.2238244.
6. Wong DWK, Liu J, Lim JH, Jia X, Yin F, Li H, Wong TY. Level-set based automatic cup-to-disc ratio determination using retinal fundus images in ARGALI. En: 30th Annual International IEEE EMBS Conference, vol. 30. 2008, p. 2266-9. https://doi.org/10.1109/IEMBS.2008.4649648.
7. Joshi GD, Sivaswamy J, Krishnadas SR. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. IEEE Trans Med Imag. 2011;30:1192–205. https://doi.org/10.1109/TMI.2011.2106509.

8. Yin F, Liu J, Wong DWK, Tan NM, Cheung C, Baskaran M, Aung T, Wong TY. Segmentación automatizada del disco óptico y la copa óptica en imágenes de fondo de ojo para el diagnóstico de glaucoma. En: 2012 25th IEEE international symposium on computer- based medical systems (CBMS). 2012, p. 1-6. https://doi.org/10.1109/CBMS.2012.6266344.
9. Cheng J, Liu J, Xu Y, Yin F, Wong DWK, Tan N-M, Tao D, Cheng C-Y, Aung T, Wong TY. Clasificación de superpíxeles basada en la segmentación del disco óptico y la copa óptica para la detección del glaucoma. En: IEEE transactions on medical imaging, vol. 32. 2013, p. 1019-32. https://doi.org/10.1109/TMI.2013.2247770.
10. Diaz-Pinto A, Morales S, Naranjo V, Alcocer P, Lanzagorta A. Glaucoma diagnosis by means of optic cup feature analysis in color fundus images. En: 24th European signal processing conference (EUSIPCO), vol. 24. 2016, p. 2055-9. https://doi.org/10.1109/EUSIPCO.2016.7760610.
11. Liu J, Zhang Z, Wong DWK, Xu Y, Yin F, Cheng J, Tan NM, Kwoh CK, Xu D, Tham YC, Aung T, Wong TY. Automatic glaucoma diagnosis through medical imaging informatics. J Am Med Inf Assoc. 2013;1:1021-7. https://doi.org/10.1136/amiajnl-2012-001336.
12. LeCun Y. Generalization and network design strategies, Technical Report CRG-TR-89-4, University of TorontoNew York: Elsevier; 1989.
13. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. Desafío de reconocimiento visual a gran escala de ImageNet. Int J Comput Vis. 2015;115(3):211-52. https://doi.org/10.1007/s11263-015-0816-y.
14. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929-58.
15. Simonyan K, Zisserman A. Redes convolucionales muy profundas para el reconocimiento de imágenes a gran escala. 2014. ArXiv e-prints arxiv:abs/1409.1556.
16. Carneiro G, Nascimento J, Bradley AP. En: Navab N, Hornegger J, Wells WM, Frangi AF, eds. Unregistered multiview mammogram analysis with pre-trained deep learning models. 2015, p. 652-60. Cham: Springer. https://doi. org/10.1007/978-3-319-24574-4_78.
17. Chen H, Ni D, Qin J, Li S, Yang X, Wang T, Heng PA. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J Biomed Health Inf. 2015;19(5):1627-36. https://doi.org/10.1109/JBHI.2015.2425041.
18. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Redes neuronales convolucionales para el análisis de imágenes médicas: ¿entrenamiento completo o ajuste fino? IEEE Trans Med Imag. 2016;35(5):1299-312. https://doi.org/10.1109/ TMI.2016.2535302.
19. Yaniv B, Idit D, Lior W, Hayit G. Aprendizaje profundo con entrenamiento no médico utilizado para la identificación de patologías torácicas. En: Proceedings of SPIE 9414, medical imaging 2015: computer-aided diagnosis. 2015, p. 94140-7. https://doi. org/10.1117/12.2083124.
20. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. 2014. arXiv e-prints arxiv:abs/1403.6382.
21. Chen X, Xu Y, Wong DWK, Wong TY, Liu J. Detección de glaucoma basada en una red neuronal convolucional profunda. En: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). 2015, p. 715-8. https://doi.org/10.1109/EMBC.2015.7318462.
22. Alghamdi HS, Tang HL, A.Waheeb S, Peto T. Automatic optic disc abnormality detection in fundus images: a deep learning approach. En: OMIA3 (MICCAI 2016). 2016, p. 17-24. https://doi.org/10.17077/omia.1042.
23. Abbas Q. Glaucoma-deep: detección de la enfermedad ocular del glaucoma en imágenes del fondo de ojo de la retina utilizando el aprendizaje profundo. Int JAdv Comput Sci Appl. 2017;8(6):41-5. https://doi.org/10.14569/IJACSA.2017.080606.
24. Orlando JI, Prokofyeva E, del Fresno M, Blaschko MB. Transferencia de redes neuronales convolucionales para la identificación automatizada del glaucoma . En: SPIE proceedings. 2017, p. 10160-10. https://doi.org/10.1117/12.2255740.
25. Budai A, Bock R, Maier A, Hornegger J, Michelson G. Robusta segmentación de vasos en imágenes de fondo de ojo. Int J Biomed Imag. 2013. https://doi.org/10.1155/2013/154860.
26. Fumero F, Alayón S, Sánchez JL, Sigut J, González-Hernández M. RIM-ONE: una base de datos de imágenes retinales abierta para la evaluación del nervio óptico. En: 2011 24th international symposium on computer-based medical systems (CBMS). 2011, p. 1-6. https://doi.org/10.1109/CBMS.2011.5999143.
27. sjchoi86: Base de datos sjchoi86-HRF. GitHub. 2017. Consultado el 2 de febrero de 2017.
28. Chollet F, et al. Keras. GitHub. 2015. Consultado el 21 de febrero de 2017.
29. Xu P, Wan C, Cheng J, Niu D, Liu J. Optic disc detection via deep learning in fundus images. Fetal, infantíl y ophthal- mic medical image analysis. Cham: Springer; 2017. p. 134-41.
30. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. En: 2015 IEEE conference on computer vision and pattern recognition (CVPR). 2015, p. 1-9. https://doi. org/10.1109/CVPR.2015.7298594.
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. En: The IEEE conference on computer vision and pattern recognition (CVPR). 2016, p. 2818-26.
32. He K, Zhang X, Ren S, Sun J Aprendizaje residual profundo para el reconocimiento de imágenes. En: The IEEE conference on computer vision and pattern recognition (CVPR). 2016.
33. Chollet F. Xception: Aprendizaje profundo con convoluciones separables en profundidad. ArXiv e-prints. 2016. arxiv :abs/1610.02357.
34. Hastie T, Tibshirani R, Friedman J. Los elementos del aprendizaje estadístico. Cham: Springer; 2009. https://doi. org/10.1007/978-0-387-84858-7.
35. Mason SJ, Graham NE. Áreas bajo las curvas de características operativas relativas (roc) y niveles operativos relativos (rol): significado estadístico e interpretación. Q J R Meteorol Soc. 2002. https://doi.org/10.1256/0035900023 20603584.
36. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 1947;18(1):50–60. https://doi.org/10.1214/aoms/1177730491.
37. Chakravarty A, Sivaswamy J Clasificación del glaucoma con una fusión de características basadas en la segmentación y la imagen. En: 2016 IEEE 13th international symposium on biomedical imaging (ISBI). 2016, p. 689-92. https://doi.org/10.1109/ ISBI.2016.7493360.