ISSN 1870-4069

Modelos de aprendizaje automático para clasificar patologías oculares en imágenes de fondo de ojo como prueba de tamizaje

Miguel Ángel Méndez Lucero[1], E. Mahuina Campos Castolo[2], Dania N. Lima Sánchez[2], Josué Fabricio Urbina González[3], Orlando Cerón Solís[2], Alejandro Alayola-Sansores[2], German Fajardo Dolci[4]

[1] University of Edinburgh,
School of Informatics, Artificial Intelligence and its Applications Institute,
Reino Unido
[2] Universidad Nacional Autónoma de México,
Facultad de Medicina, Departamento de Informática Biomédica,
México
[3] Oracle México Development Center,
México
[4] Universidad Nacional Autónoma de México,
Director de la Facultad de Medicina,
México

{Mendezluceromiguelangel, orlandoceronsolis, ale.alayola, urbgon}@gmail.com, {dibfm, ,direccionfm}@unam.mx, danianimbe@hotmail,com

Resumen. Ante la posibilidad de múltiples patologías en cada fotografía del fondo de ojo, escasez de datos para enfermedades poco comunes; y la falta de bases de datos públicas bien clasificadas para enfermedades oculares, son algunos de los principales problemas que obstaculizan la creación de clasificadores. Estos modelos podrían ayudar a los oftalmólogos en el diagnóstico de enfermedades oculares. Utilizando técnicas de transferencia de aprendizaje y aumento de datos, analizamos la eficiencia de diversos algoritmos de varias clases y métodos de conjunto. Después de comparar el rendimiento de aproximadamente 96 modelos multicategoríales, propusimos un modelo probado con 1350 imágenes; con un conjunto de datos no balanceado de 10 clases.

Palabras clave: Aprendizaje de transferencia, clasificador de enfermedades oculares, métodos de conjunto, aprendizaje profundo.

Machine Learning Models to Classify Eye Pathologies in Fundus Images as Screening Test

Abstract. The possibility of multiple pathologies in each fundus photography, scarcity of data for uncommon diseases and lack of properly classified public databases for ocular diseases are some of the main problems that hinder the

creation of classifiers. These models could assist ophthalmologists in diagnostics of eye diseases. Using transfer learning and data augmentation techniques, we analyze the efficiency of diverse multi-class algorithms and ensemble methods. After comparing the performance of approximately 96 different multi-categorical models, we proposed a model tested with 1350 images; which is an unbalanced data set of 10 classes.

Keywords: Transfer learning, eye disease classifier, ensemble methods, deep learning.

1. Introducción

En México, las discapacidades de todo tipo afectaron al 6% de la población en 2014, de las cuales el 58.7% se asociaron con discapacidad visual. Las primeras cuatro causas de discapacidad visual en México son: 1. Cataratas, 2. Retinopatía diabética, 3. Glaucoma y 4. Degeneración macular relacionada con la edad. Estas cuatro enfermedades son responsables de hasta 575,954 casos de ceguera por año y 4, 917,340 casos de otras patologías, con un costo estimado en 2013 de $ 539, 000,000 dólares estadounidenses por año [1]. Se considera que la mayor parte de las complicaciones que puede llevar a la perdida irremediable de la visión son prevenibles con diagnóstico precoz, por lo que es imprescindible tener herramientas de apoyo el diagnóstico para una detección temprana. La detección de patología por fotografía digital es ampliamente aceptada para el diagnóstico de la retina. Sin embargo, un diagnóstico definitivo requiere una evaluación completa por parte de un oftalmólogo.

En los últimos tiempos, las técnicas de aprendizaje automático se han utilizado para resolver problemas de reconocimiento y clasificación de patrones en diferentes áreas. El desarrollo de algoritmos de inteligencia artificial (IA) para resolver problemas médicos ha sido una de las áreas más prometedoras, especialmente con el fin de realizar tamizaje masivo de los pacientes, el aprendizaje automático es una rama importante de la IA, que puede dividirse en supervisado o no supervisado, cuando están las imágenes etiquetadas previamente [2]. Las técnicas de clasificación que se utilizaron son las siguientes: Clasificación mediante modelos probabilísticos: Naïve Bayes, Maquinas de Soporte vectorial (SVM por sus siglas en ingles), Bosque Aleatorio (Random Forest Regression), Regresión Logística, Naïve Bayes Multinomial.

Estos modelos son ideales cuando el número de imágenes son pequeños, para la otra alternativa, el aprendizaje profundo, se requiere regularmente cientos de miles de datos [2]. La eficacia del aprendizaje automático toma en cuenta el rendimiento del modelo, para ello se utiliza una matriz de confusión, donde la imagen propuesta será clasificada y comparada con la clasificación real, obteniendo los valores de sensibilidad y especificidad. La exactitud evalúa el entrenamiento del modelo cuando existe un amplio número de muestras, cuando existe un desbalance en la muestra de categorías es más confiable utilizar la precisión, que evalúa la calidad del modelo en la tarea de clasificación, y la métrica de exhaustividad sobre la cantidad que el modelo es capaz de identificar, es decir que porcentaje es capaz de seleccionar adecuadamente.

Para evaluar de una manera más integral este proceso, se utiliza el valor F, un valor igual a 1 considera que el modelo es perfecto y 0 implica que el modelo no se ajusta a la realidad.

Además se utiliza la curva característica operativa del receptor (COR o ROC en sus siglas en ingles). Los métodos de ensamblaje implica la combinación de múltiples modelos clasificadores, produciendo un modelo clasificador final, mientras que el aprendizaje por transferencia es un método para adaptar un modelo entrenado en un dominio a otro dominio [3].

2. Trabajos relacionados

En el campo de la medicina, es necesario tener precauciones en relación a la privacidad del paciente, lo que ocasiona que en ocasiones el acceso a grandes bases de datos públicas de patologías médicas se encuentre limitadas, esto dificulta la aplicación de aprendizaje profundo, por lo cual diversos autores han trabajado con diferentes técnicas para aumentar los datos.

En un estudio realizado por Asperti y Mastronardo, utilizaron 4000 imágenes endoscópicas de enfermedades gastrointestinales etiquetadas con médicos endoscopistas, con diferentes patologías, mostrando que la aplicación de diferentes técnicas, permitía una mejora en la clasificación [4].

En relación a las imágenes obtenidas de fondo de ojo, se han hecho múltiples investigaciones, debido a que las imagines digitales proporcionan valiosa información, no sólo con respecto a patologías, sino también para datos morfológicos que pueden analizarse de manera rápida y no invasiva utilizando inteligencia artificial, en una amplia revisión realizada por Schmidt-Erfurth y cols, con respecto a la aplicación de la IA en patologías de retina; encontraron que la mayor parte de los estudios automatizados han sido aplicados a retinopatía diabética, utilizando aprendizaje profundo para la detección, clasificación diagnóstica y orientación terapéutica en esta patología, que podría ofrecer nuevos tratamientos, y una medicina personalizada, sin embargo la mayor parte de estos avances se basan en una sola patología [5]. Desafortunadamente, la mayor parte de las patologías de fondo de ojo presentar comorbilidades, y las causas de discapacidad implican múltiples diagnósticos.

Abordando este problema, Choi y cols, utiliza el aprendizaje profundo para analizar imágenes de retina, utilizando una red neuronal convolucional mediante el uso de MatConvNet, para la detección de múltiples enfermedades de la retina, utilizando una base de datos de retina (STARE). Utilizaron 10 categorías, incluyendo retina normal y nueve enfermedades con 25 imágenes por cada grupo, obteniendo el mejor rendimiento mediante el uso de aprendizaje basado en bosques aleatorios, utilizando el modelo VGG-19. Encontraron que al aumentar el número de categorías, disminuía el rendimiento del modelo, obteniendo una precisión de 30.5% cuando incluyeron 10 categorías, una kappa de Cohen de 0.244, al contrario, cuando sólo compararon tres categorías obtuvieron una precisión de 72.8 y un valor kappa de 0.577.

El bajo rendimiento lo atribuyeron al pequeño tamaño de los conjuntos de datos, pero encontraron que al utilizar Transfer Learning pudieron mejorar el rendimiento, por lo que sugieren realizar otros estudios para confirmar la efectividad de los algoritmos [6]. Tomando estos antecedentes, decidimos evaluar un modelo con la precisión y sensibilidad suficientes para los estándares medicos, a fin de utilizarlo como prueba de detección y clasificación de enfermedades oculares.

3. Material y métodos

3.1. Obtención de datos

El conjunto de datos utilizado en este artículo fue obtenido de bases abiertas de imágenes, fotografías tomadas de un Centro de salud, y del Instituto Mexicano de Oftalmología I.A.P, aprobado por el comité de ética e investigación de esta misma institución. Se trabajó con un conjunto de 1352 imágenes, clasificados en 10 categorías verificado por un médico especialista en retina, dentro de paréntesis se menciona el número de imágenes: adelgazamiento de la retina (n=36), cataratas (n=10), drusas (n=25) ,excavación del nervio óptico/glaucoma (n=103), degeneración macular (n=15), retinopatía diabética leve (n=380), retinopatía diabética moderada (n=45), retinopatía diabética severa (n=20) , retinopatía hipertensiva (n=68) y normal (n=650). Con este banco de imágenes se hicieron cinco propuestas de clases:

1. Adelgazamiento de retina, cataratas, drusas, excavación del nervio óptico/glaucoma, degeneración macular, retinopatía diabética general, retinopatía hipertensiva y normal. (Total 8 clases).
2. Adelgazamiento de la retina, cataratas, drusas, excavación del nervio óptico/glaucoma, degeneración macular, retinopatía diabética severa, retinopatía hipertensiva y normal. (Total 8 clases).
3. Adelgazamiento de la retina, cataratas, drusas, excavación del nervio óptico/glaucoma, degeneración macular, retinopatía, retinopatía diabética moderada, retinopatía diabética severa, retinopatía hipertensiva y normal (total 10 clases).
4. Adelgazamiento de la retina, excavación del nervio óptico/glaucoma, degeneración macular, retinopatía diabética, retinopatía hipertensiva y normal (total de 6 clases).
5. Imagen de ojo normal y enfermo, conformado con las otras nueve patologías. (Total 2 clases).

En la primera subdivisión consideramos todos los diferentes grados de retinopatía diabética (retinopatía diabética leve, moderada y severa) en una clase, esta clasificación se consideró ya que formaban diferentes grados de severidad de una misma patología; por otro lado, para la segunda subdivisión consideramos sólo la retinopatía diabética severa, dado que tenía diferencias importantes en sus alteraciones, con el resto de las imágenes. Para la cuarta subdivisión incluimos la categoría de Drusas como parte de degeneración macular, ya que comparten la misma presentación clínica y excluimos cataratas debido a la falta de datos.

3.2. Aumento de datos

Se ha demostrado que el uso de técnicas de aumento de datos en conjuntos que tienen una muestra pequeña con clases desequilibradas, mejora significativamente los puntajes de precisión, exhaustividad y el valor Fde los modelos entrenados [4].

Para el aumento de datos utilizamos los siguientes parámetros: rotation_range=180°, width_shift_range=12%, height_shift_range=12%, zoom_range=12%, shear_range= 20%, horizontal_flip=True, fill_mode='nearest'.

Además utilizamos dos enfoques. El primero fue aplicar el aumento de datos para igualar las demás clases a la clase con el mayor número de datos (la clase de ojo normal, que contenía 650 imágenes), con lo cual al final del proceso teníamos 650 imágenes para cada clase. El segundo enfoque fue reducir el número de imágenes para cada clase, de modo que estuvieran equilibradas. En este caso, redujimos el número de imágenes por clase a 32. Luego, el proceso de aumento de datos genero nueve veces más imágenes, dándonos un total de 320 imágenes por clase. Finalmente, analizamos el rendimiento, entrenando varios modelos usando estas subdivisiones de clases.

3.3. Transfer learning con aprendizaje profundo

Las técnicas de aprendizaje profundo evolucionan constantemente debido a su gran uso en las tareas de clasificación [7], además las redes neuronales convolucionales han demostrado sobresalir en tareas que involucran la clasificación de imágenes [8]. En el caso de nuestro artículo, las técnicas de transferencia de aprendizaje se usan como un extractor de funciones [9]. Debido a la falta de datos en nuestra población problema y la gran diferencia entre las imágenes utilizadas para entrenar la CNN contra las contenidas en nuestro conjunto de datos. Utilizamos la CNN pre entrenada para extraer características relevantes de cada imagen [10] y luego alimentarlas a un algoritmo más simple de múltiples clases.

Si bien hay muchos CNN pre entrenado que se pueden usar para la extracción de características, en este enfoque utilizamos el modelo VGG16, debido a su estructura relativamente simple (16 capas) y al buen rendimiento. Este modelo ganó el ImageNet ILSVRC-2014, también vale la pena mencionar que debido a la simplicidad de su arquitectura, el proceso de la función de extracción tarda 55 segundos para aproximadamente 7000 imágenes.

3.4. Algoritmos de clases múltiples y métodos de ensamble

Para encontrar el modelo que mejor se ajustara y clasificara nuestros datos pre procesados, usamos e implementamos varios algoritmos de clases múltiples. Primero, sin ningún método de ensamblaje, implementamos algoritmos simples de varias clases, tales como: Máquina de Soporte Vectorial (SVM), Bosques aleatorios (con una profundidad de 8 niveles), Regresión logística, Bernoulli Naïve Bayes y Multinomial Naïve Bayes. Además, implementamos varios métodos de ensamble [11], como Bagging Classifier utilizando los algoritmos mencionados anteriormente como estimadores base y bosque aleatorio. Éste clasificador es un meta estimador de ensamblaje que combina varios algoritmos de aprendizaje automático y los ajusta con subconjuntos aleatorios del conjunto de datos de entrenamiento. Al igual que Bosques aleatorios, los clasificadores de ensamblaje funcionan como métodos de promedio. Esto significa que usamos cada uno de los modelos ajustados y promediamos sus predicciones que resultan en un estimador combinado que generalmente es mejor que cualquiera de los estimadores de base debido a que su varianza se reduce.

3.5. Prueba de conjunto de datos y métricas para evaluar el rendimiento

Después de entrenar nuestros modelos, el departamento de informática biomédica recopiló un conjunto de datos de prueba que consta de un conjunto de 121 dividido en las 10 clases iniciales para evaluar el rendimiento de cada modelo entrenado. En los datos utilizados para la prueba, se detectó un pequeño sesgo debido a un diseño de muestra no óptimo causado por la disponibilidad de datos y tiempo; que se ajustó para poder reproducir de manera más confiable el rendimiento real de cada modelo.

Para evaluar el rendimiento de cada modelo entrenado, utilizamos diferentes medidas para la clasificación [12].

Puntuación de precisión: esta función calcula la precisión de las predicciones correctas:

$$\frac{\sum_{i=1}^{c} TP_i}{n_{muestra}}. \tag{1}$$

Simbología: la variable 'c' representa el número de clases, tasa de verdaderos positivos (TP), tasa de verdaderos negativos (TN), tasa de falsos positivos (FP), tasa de falsos negativos (FN); El hiperparámetro ß es igual a uno, ya que estamos calculando la puntuación de F1.

Exhaustividad: efectividad promedio por clase de un clasificador para identificar etiquetas de clase:

$$\frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i + FP_i}}{c}. \tag{2}$$

Precisión: un acuerdo promedio por clase de las etiquetas de clase de datos con las de un clasificador:

$$\frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i + FN_i}}{c}. \tag{3}$$

Puntuación F1: relación entre las etiquetas positivas de los datos y las dadas por un clasificador basado en un promedio por clase:

$$\frac{(\beta^2 + 1)Sensibilidad \times Precisión}{\beta^2 Sensibilidad + Precisión}. \tag{4}$$

ß es igual a uno para calcular la puntuación de F1.

4. Resultados

4.1. Enfoque de aumento de datos

Se utilizaron varios algoritmos de clases múltiples, utilizando dos enfoques diferentes de aumento de datos. Para medir el rendimiento de estos enfoques, utilizamos las primeras tres subdivisiones de clase; usando la puntuación de precisión para medir ambos enfoques.

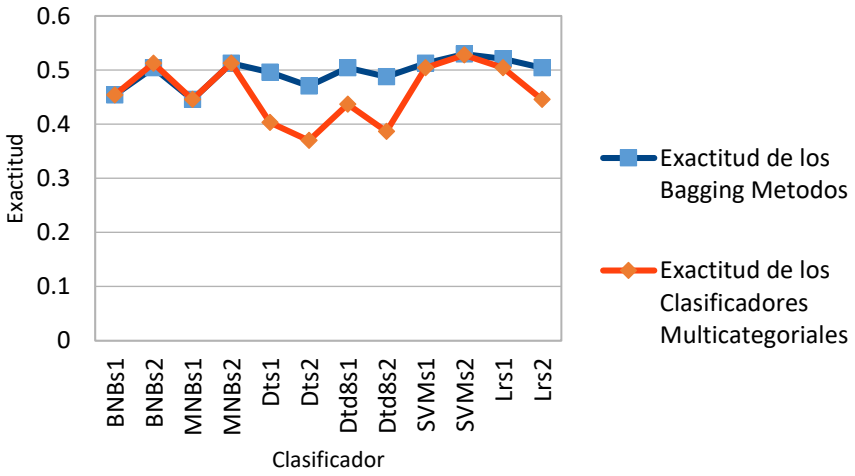

Fig. 1. Método Bagging Vs. Clasificadores Multicategoríales.

Observamos que el primer enfoque sobresalió sobre el segundo en casi todos los clasificadores en un rango entre 5% y 25%, con la excepción de Multinomial y Bernoulli Naïve Bayes, en el que el segundo enfoque sobresalió sobre el primero con 5% más de precisión.

4.2. Algoritmos de clases múltiples frente a métodos Bagging

Usando las dos primeras subdivisiones, medimos la eficiencia que los métodos de Ensemble realizaron en algoritmos normales de clases múltiples Figura 1. Los métodos de embolsado mejoran la puntuación de precisión en un rango entre 2% -8%, dependiendo del algoritmo de clasificación.

Vale la pena mencionar que aunque el número de estimadores utilizados para el entrenamiento fue bajo (10 estimadores de línea de base), un análisis posterior mostró que con 100 estimadores de base en algunos modelos, la precisión mejoró en un 2% (el caso de regresores lineales y decisión árboles), mientras que los demás mantuvieron su precisión sin cambios.

Este gráfico representa el rendimiento de cada modelo de múltiples categorías en comparación con su implementación como un estimador de referencia en el método de ensacado. Los modelos de bagging en este cuadro se construyeron utilizando estimadores de 10 bases y permitiendo un arranque.

Los clasificadores multicategoría utilizados en este cuadro son: Máquina de vectores de soporte (SVM), Árbol de decisiones (Dt y Dtd8, donde d8 representa la profundidad máxima utilizada en el árbol de decisión), Regresión logística (LR), Bayes ingenuos de Bernoulli (BNB), Bayes ingenuos multinomiales (MNB). El subfix s1 y s2 representan la subdivisión de la categoría utilizada para el entrenamiento.

Tabla 1. Resultados de rendimiento de modelos de aprendizaje profundo de múltiples categorías para cada subdivisión de clase.

Numero de Categorías	Exactitud	Sensibilidad	Precisión	F
10 Clases	0.4308	0.4332	0.2012	0.1773
8 Clases	0.5294	0.5629	0.2503	0.2383
6 Clases	0.5855	0.6003	0.3380	0.3303
2 Clases	0.8884	0.8973	0.8915	0.8943

4.3. Conjuntos de subdivisión de clase y análisis de estimador base

Como se esperaba, el rendimiento de cada subdivisión aumentó en proporción al número de clases, la Tabla 1 muestra el modelo con el mejor rendimiento en cada subdivisión. Encontramos un aumento general del rendimiento en la subdivisión 2 sobre la subdivisión 1 en un rango entre 4% y 16% (ambos midieron las mismas clases 6 y 8). Por lo tanto, la Tabla 1 solo se muestra los mejores resultados obtenidos de la subdivisión 2 en la clasificación de seis y ocho categorías.

Además, encontramos que los resultados sobresalieron en desempeño en comparación con investigaciones similares [6]; utilizando un conjunto de datos de entrenamiento más pequeño (después de aplicar técnicas de aumento de datos) y para probar, un conjunto de datos diferente para eliminar el sesgo y evaluar de manera más confiable el rendimiento de cada modelo.

En la Tabla 1 podemos observar los resultados de mejor rendimiento de modelos de aprendizaje profundo de múltiples categorías para cada subdivisión de clase. En todas las subdivisiones, la máquina de vectores de soporte de ensacado fue el modelo con los mejores puntajes de rendimiento. Los modelos con 8 clases fueron entrenados por las subdivisiones 1 y 2; Los resultados que se muestran aquí se obtuvieron del modelo entrenado por la segunda subdivisión. En la clasificación múltiple, el mejor rendimiento se obtuvo en la subdivisión de 6 clases, en la que dos modelos superaron a los demás.

El primero fue entrenado con Bayes Naïve multinomial como estimador base, obteniendo una sensibilidad del 61.38%, una precisión del 56.75% y una puntuación F1 del 35.85%, mientras que la segunda máquina de vectores de soporte obtuvo una sensibilidad o 61.38%, una precisión de 56.75% y un F1 - Puntaje de 35.85%. De estos resultados podemos concluir que los modelos probabilísticos son tan buenos estimadores como los modelos de regresión que normalmente se destacan cuando se combinan con bosques aleatorios en este tipo de tareas.

Para el problema de la clasificación múltiple como se muestra en la Figura 2, el mejor rendimiento se obtuvo en la subdivisión de 6 clases, en las que dos modelos superaron a los demás. El primero fue entre sensibilidad del 61,38%, una precisión del 56,75% y una puntuación F1 del 35,85%, mientras que la segunda máquina de vectores de soporte obtuvo una sensibilidad del 61,38%, una precisión del 56,75% y una puntuación F1 del 35,85%.

En la Figura 2 observamos los resultados de las métricas utilizadas para evaluar el rendimiento de varios estimadores básicos en un problema de clasificación de 6 clases.

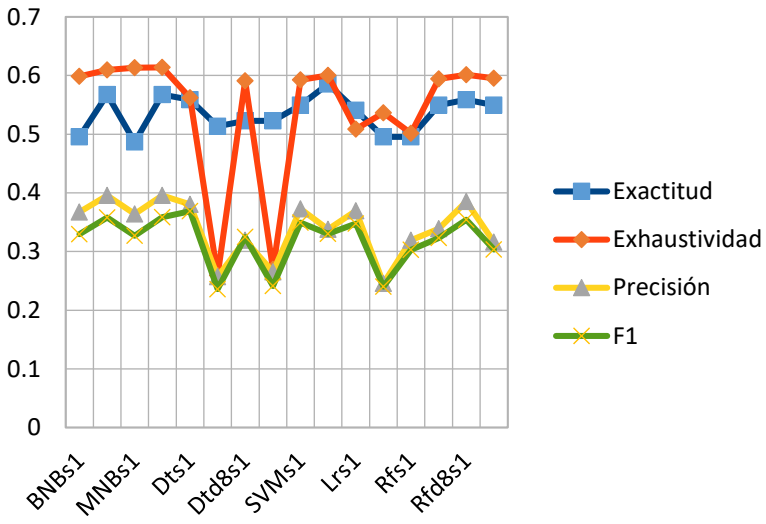

Fig. 1. Medidas de Rendimiento de los métodos Bagging de seis clases.

Los estimadores básicos utilizados en este gráfico son la máquina de vectores de soporte (SVM), el árbol de decisión (Dt y Dtd8 donde d8 representa la profundidad máxima utilizada en el árbol de decisión), la regresión logística (LR), las bayas ingenuas de Bernoulli (BNB), Bayes ingenuos multinomiales (MNB) y bosque aleatorio (RF). El subfijo s1 representa los modelos que usaron todos los niveles de retinopatía diabética para entrenar a la clase de retinopatía diabética, mientras que s2 representa los modelos que usaron solo la retinopatía diabética severa para entrenar a la clase de retinopatía diabética.

Podemos concluir con estos resultados que los modelos probabilísticos son tan buenos estimadores como los modelos de regresión, que normalmente se destacan junto con bosques aleatorios en este tipo de tareas. Para el algoritmo de prueba de detección se realizó una ligera modificación de los datos para medir el rendimiento. Para el conjunto de datos inicial tuvimos 650 que pertenecían a la clase normal y 702 en total para las patologías. Dividimos el conjunto de datos en 90% para capacitación y 10% para pruebas, luego aplicamos el aumento de datos para duplicar la cantidad de datos para capacitación.

Se utilizaron varios algoritmos para entrenar los modelos: SVM, Bayes Naïve de Bernoulli, Bayes Naives multinomiales (MBN), árboles de decisión y regresión logística. El algoritmo SVM fue mejor sobre el segundo en casi todas las subclasificaciones (entre 5% y 25%), con la excepción de los basados en Bayes, que pudieron clasificar un 5% más preciso la segunda propuesta de subclasificación de clases.

Usando las dos primeras subdivisiones, medimos la eficiencia de los métodos de conjunto en algoritmos normales de clases múltiples (Figura 1). Los métodos de ensamble mejoran el puntaje de precisión en un rango de 2% -8%, dependiendo del algoritmo de clasificación. Vale la pena mencionar que, a pesar de que el número de

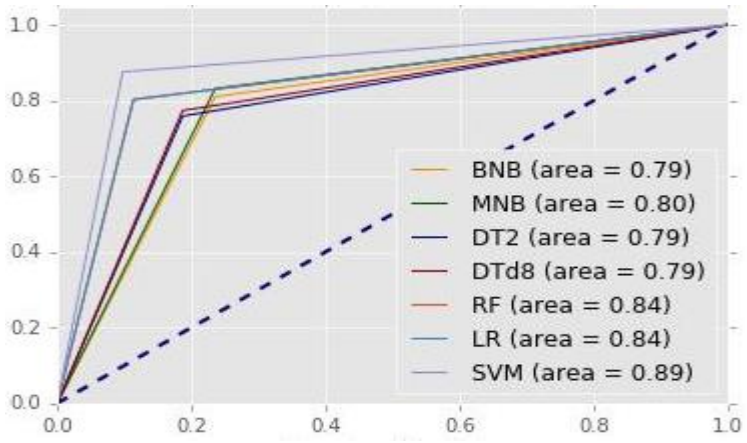

Fig. 3. Curvas características de funcionamiento del receptor de varios modelos entrenados.

estimadores utilizados para el entrenamiento es bajo (10 estimadores de referencia), un análisis posterior mostró que, con 100 estimadores básicos en algunos modelos, la precisión mejoró en un 2% (este fue el caso de las regresiones lineales y el árboles de decisión), mientras que los demás mantuvieron su precisión sin cambios.

Los modelos fueron entrenados utilizando las características extraídas de un conjunto de datos de imágenes de fondo con el VGG-16. El conjunto de datos se dividió en 90% para el entrenamiento y 10% para las pruebas. Las técnicas de aumento de datos se aplicaron al conjunto de datos de entrenamiento para mejorar la sensibilidad de los modelos. Para el algoritmo de prueba de detección, se realizó una ligera modificación de los datos para medir el rendimiento. Del conjunto de datos inicial, tuvimos 650 que pertenecían a la clase normal y 702 en total para las patologías. Dividimos el conjunto de datos en 90% para entrenamiento y 10% para pruebas, luego aplicamos el aumento de datos para duplicar la cantidad de datos para entrenamiento.

Los resultados del rendimiento se describen en la curva ROC que se muestra en la Figura 3. Una vez más se revela que SVM es el algoritmo que destaca sobre el resto de las técnicas de aprendizaje automático, mostrando una sensibilidad del 89,7%, seguido de RF y LR. Los resultados del rendimiento se describen en la curva ROC que se muestra en la Figura 3. Una vez más se revela que SVM es el algoritmo que destaca sobre el resto de las técnicas de aprendizaje automático, mostrando una sensibilidad del 89,7%, seguido de RF y LR.

En la Figura 3 se puede observar las Curvas ROC como prueba de detección para las enfermedades en imágenes de fondo de ojo.

Los modelos fueron entrenados utilizando las características extraídas de un conjunto de datos de imágenes de fondo con el VGG-16. El conjunto de datos se dividió el 90% para el entrenamiento y el 10% para las pruebas, se aplicaron técnicas de aumento de datos al conjunto de datos del entrenamiento para mejorar la sensibilidad de los modelos.

5. Conclusiones y trabajo a futuro

Se realizó un análisis en varias subdivisiones de clase para aumentar la puntuación de precisión de los modelos; cada subdivisión fue creada cuidadosamente para evaluar el desempeño de varias clases. Para la primera y segunda subdivisiones, analizamos el efecto de usar solo imágenes de retinopatía severa durante el entrenamiento contra la retinopatía general (todos los grados de retinopatía). Observamos que, con la excepción de los modelos entrenados con MNB, los resultados mostraron que cuando se utilizó únicamente retinopatía severa, tuvo mejor rendimiento. Confirmando que la razón principal es probablemente las características bien definidas mostradas en la retinopatía severa contra los otros niveles de retinopatía

El segundo análisis mostró que las cataratas, drusas y retinopatía hipertensiva son las patologías con menos precisión en el rendimiento general del clasificador. Las cataratas se excluyeron en la subdivisión 4 ya que los resultados mostraron que la falta significativa de datos está altamente relacionada con el bajo rendimiento de los modelos que predecían esta clase. Es probable que para la evaluación de cataratas sea necesario un modelado propio.

Otra de las dificultades en la clasificación fue en degeneración macular relacionada con la edad (DMAE), ya que se menciona que puede contener drusas [13]. Si bien las drusas no se considera una causa directa de degeneración macular, es uno de los síntomas visibles observados en las imágenes de fondo de ojo que contienen etapas iniciales e intermedias de DMAE. Debido a esto, la subdivisión cuatro incluyó datos de drusas como parte de la clase DMAE. Los resultados de la prueba mostraron una disminución en el rendimiento de los modelos que incluyen Drusas como parte de DMAE, en comparación con aquellos entrenados con solo DMAE y un probable factor confusor es que se incluyó todos los tipos de DMAE (es decir, degeneración macular seca y húmeda en todos sus grados). Con el fin de eliminar estos sesgos, será deseable que en estudios futuros se tome en cuenta estas subdivisiones de la enfermedad.

La investigación actual en inteligencia artificial aplicados a imágenes de retina, generalmente involucra grandes conjuntos de datos, con cientos de miles de imágenes para capacitación y validación, lo que facilita la obtención de mejores resultados debido a la cantidad de datos discriminatorios disponibles para la toma de decisiones [14], sin embargo, en nuestra población no resulta tan accesible tener este número de imágenes, por lo cual, aunque el rendimiento general del algoritmo de varias clases (es decir, el algoritmo clasificador de 6 clases) no es adecuado para ser considerado para un diagnóstico confiable de la enfermedad, tiene un mejor rendimiento comparado con un trabajo similar [6].

Como se mencionó anteriormente, la investigación se ha llevado a cabo en el campo de los algoritmos de aprendizaje automático utilizados en la tarea de clasificar patologías oculares, pero solo en problemas específicos, como la identificación de una patología específica [14]. La investigación generalmente involucra grandes conjuntos de datos, con cientos de miles de imágenes para capacitación y validación, lo que facilita la obtención de mejores resultados debido a la cantidad de datos discriminatorios disponibles para la toma de decisiones. Por esta razón, si bien el rendimiento general del algoritmo de varias clases (es decir, el algoritmo clasificador de 6 clases) no es adecuado para ser considerado para un diagnóstico confiable de la enfermedad, pero tenían un mejor rendimiento, en comparación con un trabajo similar

[6], a pesar de usar un conjunto de datos de entrenamiento más pequeño (después de aplicar técnicas de aumento de datos).

Esperábamos un valor de más del 95% de exactitud para considerarlo óptimo para uso médico, sin embargo el valor obtenido sigue siendo significativo en comparación con los resultados en otras publicaciones, donde los algoritmos tienen una sensibilidad del 80% para varias patologías [6] o, en el caso de la prueba de detección, 97% con el uso de cientos de miles de imágenes [14].

En el caso de los resultados del algoritmo de detección, obtuvimos valores prometedores al utilizar la detección de 10 patologías, ya que existen reportes para siete patologías. Se encontraron varias limitaciones y dificultades durante la realización de este trabajo; el principal fue la compilación del conjunto de datos, así como que las diferentes categorías de enfermedad, tendrían diferentes valores de muestra, además de que varios pacientes presentaban múltiples patologías, que es bastante común en el área médica. Este problema generó un sesgo en nuestro conjunto de datos que puede explicar los resultados al clasificar múltiples patologías.

Los resultados que se muestran tanto en el algoritmo de prueba de detección, como en el algoritmo de clasificación múltiple nos permiten desarrollar nuevos modelos con un conjunto de datos más amplio que podría ser utilizado por el personal de atención sanitaria, enseñanza y para realizar clasificación de otro tipo de imágenes.

Referencias

1. López-Star, E.M., Allison-Eckert, K., Limburg, H., Brea-Rodríguez, I., Lansingh, V. C.: Evaluación rápida de la ceguera evitable, incluida la retinopatía diabética, en Querétaro, México. Revista Mexicana de Oftalmología, 92, pp. 84–93 (2018)
2. Tong, T., Lu, W., Yu, Y., Shen, Y.: Application of machine learning in ophthalmic imaging modalities. Eye Vis (Lond) 16(22) (2020)
3. Lim, G., Bellemo, V., Xie, Y., Lee, X.Q., Yip, M.Y.T., Ting, D.S.W.: Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. Eye Vis (Lond), pp.14–21 (2020)
4. Asperti, A., Mastronardo, C.: The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images. In: S. Wiebe, H. Gamboa, A. Fred, & S. Bermúdez i Badia (Eds.), Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, 2: KALSIMIS. Madeira Portugal, pp. 199–205 (2018)
5. Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunović, H.: Artificial intelligence in retina. Progress in Retinal and Eye Research, 67, pp. 1–29 (2018)
6. Choi, J.Y., Yoo, T.K., Seo, J.G., Kwak, J., Um, T.T., Rim, T.H.: Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PloS ONE, 12, pp. 1–16 (2017)
7. Torrey, L., Shavlik, J.: Transfer learning. In: E. Soria, J. Martin, R. Magdalena, M. Martinez & A. Serrano, editor, Handbook of Research on Machine Learning Applications. IGI Global. pp 1–22 (2009)
8. Waseem, R.: Deep convolutional neural networks for image classification: a comprehensive review. 2449, pp. 2352–2449 (2017)
9. Garcia-Gasulla, D., Vilalta, A., Ayguad, E., Cort, U.: On the behaviour of convolutional nets for feature extraction, pp. 1–29 (2018)
10. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In Advances in neural information processing systems, pp. 3320–3328 (2014)

11. Dietterich, T.: Ensemble methods in machine learning. in international workshop on multiple classifier systems, pp. 1–15 (2000)
12. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45, pp. 427–437 (2009)
13. Bowes-Rickman, C., Farsiu, S., Toth, C.A., Klingeborn, M.: Dry age-related macular degeneration: mechanisms, therapeutic targets, and imaging. Investigative Ophthalmology and Visual Science, 54, pp. 68–80 (2013)
14. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Webster, D. R.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22) pp. 2402–2410 (2016)