

Procesamiento Automático de Imágenes Médicas para Identificar Cáncer De Piel

Víctor Manuel Castellano Ornelas¹, Gustavo Rafael Guzman Loreto², Héctor Aguilar Galván¹, André Ramos Moreno² y Rafael Guzmán Cabrera^{1,2},
Universidad de Guanajuato, Campus Irapuato Salamanca, División de Ingenierías¹, Campus León, División de ciencias e ingenierías²

Resumen

Cada vez es más común el uso de sistemas basados en inteligencia artificial para procesar grandes volúmenes de imágenes médicas de manera automática y eficiente. La finalidad es fortalecer el diagnóstico médico sin depender de un experto humano que evalúe las imágenes bajo estudio una por una. En el presente trabajo se muestran resultados de llevar a cabo la clasificación de imágenes médicas correspondientes a pacientes con cáncer de piel. Las imágenes utilizadas son de la base de datos HAM-10000, que contiene diez mil imágenes con siete tipos de cáncer de piel. La finalidad de este trabajo es explorar distintos escenarios y métodos de clasificación para determinar cuáles presentan un mejor desempeño en las métricas de evaluación utilizadas.

Palabras clave: clasificador; cáncer de piel; algoritmos de clasificación; imágenes médicas.

Introducción

Según datos de World Cancer Research Found, el cáncer de piel es el décimo noveno tipo de cáncer más común en el mundo, a su vez que el cáncer de piel no melanocítico se posiciona como el quinto más común, con 300,000 y 1,000,000 de casos respectivamente solo en el año 2018 [1], con lo anterior se pone en perspectiva el problema del cáncer de piel en el mundo, sin embargo, el verdadero riesgo es la subestimación de este fenómeno, pues a pesar de las cifras mencionadas, muchos países ni siquiera llevan un registro ordenado de casos y tratamientos de cáncer de piel, lo que claramente habla de la poca importancia que se da a este problema al menos en perspectiva.

Por medio de las herramientas informáticas utilizadas para procesar imágenes de manera automática, en el presente trabajo se propone una solución ágil a la detección de este padecimiento, utilizando la inteligencia artificial y distintos métodos de aprendizaje automático que permitan captar las características de las imágenes que eventualmente permitan al sistema implementado distinguir una clase de otra, es decir imágenes que correspondan a pacientes con cáncer de piel de imágenes correspondientes a pacientes sanos, primeramente. Y en un paso posterior distinguir entre los distintos tipos de cáncer de piel bajo estudio.

Al emplear el término Inteligencia Artificial (IA) resulta evidente que se trata de la capacidad de un ente para entender, razonar o interpretar, claro está que, dotado de una naturaleza artificial, valga la redundancia, puesto en términos menos ambiguos, la IA es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano [3]. Otro término que corresponde clarificar aquí es el de procesamiento de imágenes, que para el caso sería equivalente en definición a procesamiento de imágenes digitales, esto es, una serie de técnicas que comprenden operaciones cuyo origen es una imagen y cuyo resultado final es otra imagen, el objetivo de estas técnicas es procesar una imagen de tal modo que la resultante sea más adecuada que la imagen original para una aplicación específica [5]. Teniendo este precedente, se entiende el procesamiento de una imagen médica como la serie de técnicas necesarias aplicadas a una imagen de una parte específica del cuerpo humano que tienen por objetivo hacer más evidente (por medio de una imagen de salida) determinadas características para detección, estudio y prevención de enfermedades.

Metodología

A continuación, apoyado en el diagrama de la Figura 1 se explica el proceso que se llevó a cabo, desde la base de datos con las imágenes a trabajar, hasta la comparación de resultados, posteriormente se explica cada elemento relevante del proceso.

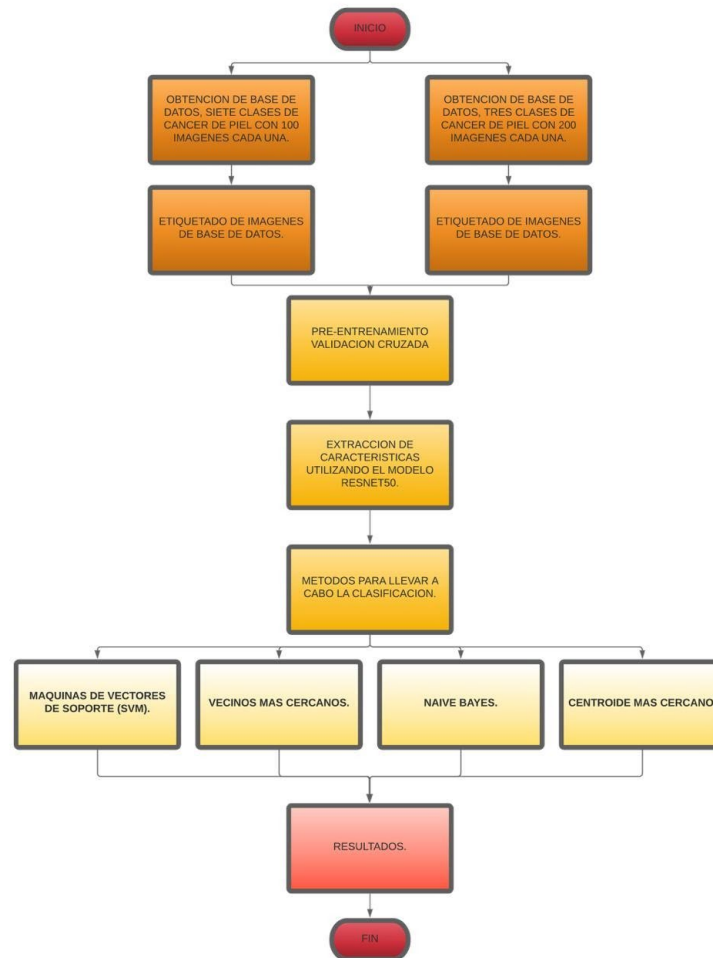


Figura 1. Diagrama de flujo del procesamiento de las imágenes médicas.

Algoritmos

- i. *Máquinas de Vectores de Soporte (SVM)*: Son un conjunto de algoritmos de aprendizaje supervisado. Para el clasificador construido se utiliza el preentrenamiento de validación cruzada y se obtienen los parámetros estadísticos a medir por medio de los siguientes algoritmos que serán descritos a continuación. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase. [4]

- ii. *Vecinos más cercanos (VMC)*: Análisis de vecinos más próximos es un método para clasificar casos basándose en su parecido a otros casos. En el aprendizaje automático, se desarrolló como una forma de reconocer patrones de datos sin la necesidad de una coincidencia exacta con patrones o casos almacenados. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Por lo tanto, la distancia entre dos casos es una medida de disimilitud. Los casos próximos entre sí se denominan “vecinos”. Cuando se presenta un nuevo caso (reserva), se calcula su distancia con respecto a los casos del modelo. Las clasificaciones de los casos más parecidos (los vecinos más próximos) se cuadrán y el nuevo caso se incluye en la categoría que contiene el mayor número de vecinos más próximos. [2]
- iii. *Naive Bayes (NB)*: Tanto en probabilidad como en minería de datos, un clasificador ingenuo Bayesiano es un método probabilístico que tiene sus bases en el teorema de Bayes y recibe el apelativo de ingenuo dadas algunas simplificaciones adicionales que determinan la hipótesis de independencia de las variables predictoras. En términos sencillos, el clasificador ingenuo de Bayes asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica. Un clasificador ingenuo de Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características. [7]
- iv. *Centroide más cercano (CMC)*: El algoritmo de clustering más usado es K-Means. Tiene una muy buena escalabilidad con la cantidad de datos. Para utilizar K-Means debemos especificar el número de grupos que queremos encontrar. A este número de grupos se le denomina K. El algoritmo K-Means sigue los siguientes pasos:
Inicialización: se elige la localización de los centroides de los K grupos aleatoriamente
Asignación: se asigna cada dato al centroide más cercano.
Actualización: se actualiza la posición del centroide a la media aritmética de las posiciones de los datos asignados al grupo.
El segundo y tercer paso se siguen iterativamente hasta que no haya más cambios. [8]

Base de datos

Para este trabajo se implementó un clasificador de siete clases (en primera instancia al menos) y posteriormente se interioriza a únicamente tres clases, con fines de mejoras en la precisión, para el caso, las imágenes con las que este clasificador trabaja fueron previamente extraídas de la base de datos HAM10000 y manualmente clasificadas para el entrenamiento. En la base de datos anterior se trabajan las siguientes clases, que representan cada una a un tipo diferente de cáncer de piel; carcinoma de células basales (bcc), queratosis benigna (bkl), dermatofibroma (df), nevos melanocíticos (nv), melanoma (mel) y lesiones de piel vasculares (vasc).

Se conoce que la mayor parte de estas imágenes han sido confirmadas y etiquetadas por expertos y el resto se trata de casos con seguimiento médico, por parte de dos institutos especializados.

Para la primer parte del experimento, se extrajeron 100 imágenes de cada clase antes mencionada, y para dar seguimiento y ver la respuesta del clasificador, se dispuso a trabajar en una segunda fase con un marco de referencia diferente, aunque igualmente basado en la HAM10000 [6], pero esta vez únicamente con tres diferentes clases (bkl, mel y nv) y con 200 imágenes de cada clase, para acentuar la mejora de los resultados, que serán comparados a continuación.

Resultados

Para un análisis cuantitativo de los resultados para un proceso de clasificación de esta naturaleza es común utilizar los siguientes parámetros; precisión, exactitud, exhaustividad y F1, los cuales se definen en términos de Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN).

$$\text{Exactitud} = \frac{VP}{VP+FP} \quad (1)$$

$$\text{Precisión} = \frac{VP+VN}{VP+FP+VN+FN} \quad (2)$$

$$\text{Exhaustividad} = \frac{VP}{VP+FN} \quad (3)$$

$$F1 = \frac{2VP}{2VP+FN+FP} \quad (4)$$

Habiendo definido los parámetros para comparar los cuatro algoritmos implementados, se presenta en la Tabla 1 y Tabla 2 los resultados para las dos bases de datos propuestas.

Tabla 1. Primer base de datos; siete clases, 100 imágenes por clase.

Parámetro	SVC	Vecinos Cercanos	Naive Bayes	Centroide Cercano
Precisión	0.56 ± 0.05	0.48 ± 0.06	0.46 ± 0.04	0.47 ± 0.02
Exactitud	0.57 ± 0.05	0.52 ± 0.07	0.48 ± 0.06	0.48 ± 0.04
Exhaustividad	0.56 ± 0.05	0.48 ± 0.06	0.46 ± 0.04	0.47 ± 0.02
F1	0.56 ± 0.05	0.48 ± 0.06	0.45 ± 0.05	0.46 ± 0.03

Tabla 2. Segunda base de datos; tres clases, 200 imágenes por clase.

Parámetro	SVC	Vecinos Cercanos	Naive Bayes	Centroide Cercano
Precisión	0.73 ± 0.04	0.65 ± 0.05	0.66 ± 0.05	0.65 ± 0.05
Exactitud	0.75 ± 0.05	0.66 ± 0.05	0.66 ± 0.05	0.65 ± 0.05
Exhaustividad	0.73 ± 0.04	0.65 ± 0.05	0.66 ± 0.05	0.65 ± 0.05
F1	0.73 ± 0.05	0.65 ± 0.05	0.66 ± 0.05	0.64 ± 0.05

A partir de este punto, es claramente observable que todas las métricas aumentaron de manera considerable, esto es debido al aumento de imágenes por clase (de 100 a 200) que se realizó, con lo que podemos comprobar la tesis de que los métodos de aprendizaje son más eficientes mientras se les entrene con un mayor número de imágenes médicas (como es el caso).

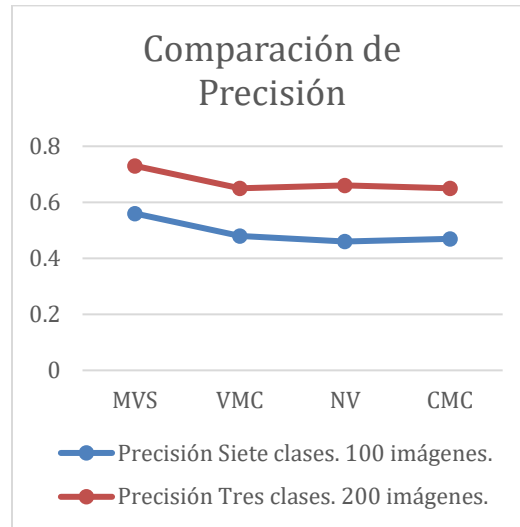


Figura 2. Comparación de los resultados del parámetro Precisión.

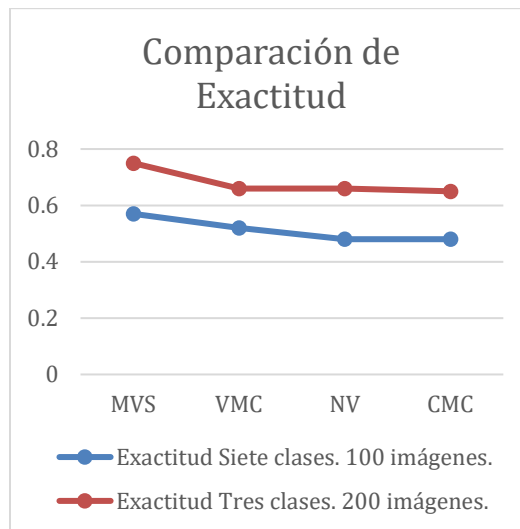


Figura 3. Comparación de los resultados del parámetro Exactitud.

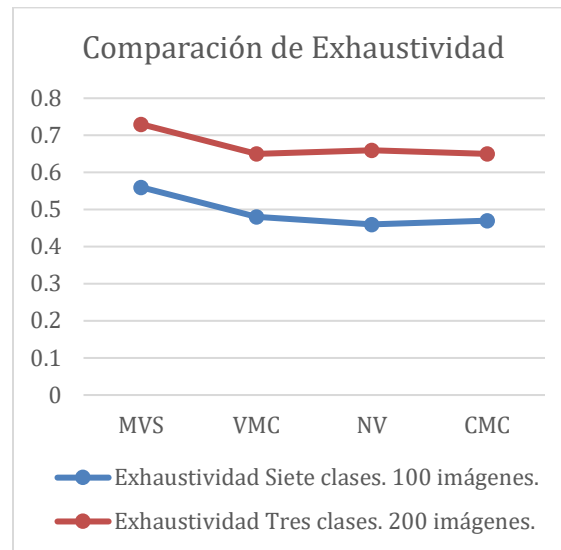


Figura 4. Comparación de los resultados del parámetro Exhaustividad.

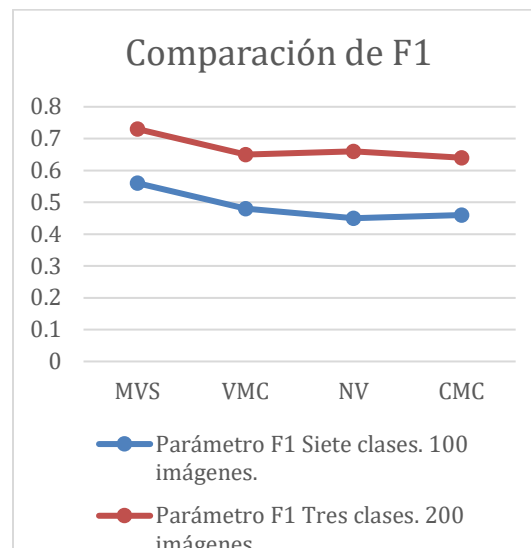


Figura 5. Comparación de los resultados del parámetro F1.

La idea de la Figura 2, Figura 3, Figura 4 y Figura 5 es mostrar la diferencia y mejora, de cada uno de los parámetros ya conocidos para medir los diferentes algoritmos utilizados para clasificar con un enfoque comparativo entre los dos escenarios propuestos.

En este apartado es importante señalar que para cada resultado mostrado en las gráficas hay un margen de variación de 0.05 en promedio para la mayoría (se puede consultar en Tabla 1 y Tabla 2), además, otro aspecto relevante a resaltar es que la mejora para el segundo escenario de clasificación (200 imágenes para cada clase, con tres clases) tienen una variación muy parecida para cada uno de los parámetros.

Conclusión

Con base en los resultados mostrados en el apartado anterior se puede concluir que al realizar acciones como aumentar el número de imágenes para una base de datos y reducir el número de clases puede aumentar significativamente parámetros importantes; Precisión, Exactitud y Exhaustividad, lo cual, siguiendo el hilo del estudio será adecuado, utilizando los mismos datos, se puede contar con un sistema que permita detectar de manera sumamente eficiente y de manera automática y rápida un mal como el cáncer de piel, incluso de manera remota si se implementa un sistema web y la imagen es cargada a distancia. Permitiendo de esta manera no solo la ayuda en la detección, sino, además, el seguimiento en la evaluación de la lesión en el tiempo.

Pasando a la parte de los algoritmos implementados en los dos escenarios de clasificación, los resultados muestran también que para todos los casos la clasificación por Maquinas de Vectores de Soporte es la que arroja parámetros más elevados, por lo tanto, es el mejor para el caso.

De manera concluyente, utilizar el algoritmo de Maquinas de Vectores de Soporte para un escenario de clasificación con únicamente tres clases y 200 imágenes para cada clase, fue el caso que mejores resultados arrojó para este estudio particular, superando en parámetros de Precisión, Exactitud, Exhaustividad y F1 a los demás casos, teniendo resultados cercanos al 80%, lo que para el caso es relevante y prueba que las técnicas utilizadas mejoran los resultados de clasificación.

Bibliografía/Referencias

- [1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, in press.
- [2] Análisis vecino más cercano: <https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=features-nearest-neighbor-analysis>. (Fecha de acceso: 01/07/2021).
- [3] Lasse Rouhiainen. (2018). *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro futuro*. Barcelona, España: Editorial Planeta.
- [4] Enrique J. Carmona Suárez, (2016, Nov), Tutorial sobre Máquinas de Vectores Soporte (SVM), Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED).
- [5] C. Pinilla, A. Alcalá y F. J. Araiza. (1997). Filtrado de imágenes el dominio de la frecuencia. Departamento de Ingeniería Cartográfica y Fotogramétrica. Universidad de Jaén.
- [6] Tschandl, P. Harvard Dataverse <https://doi.org/10.7910/DVN/DBW86T> (2018).
- [7] Naive Bayes – Teoría: <https://aprendeia.com/naive-bayes-teoria-machine-learning/>. (Fecha de acceso: 01/07/2021).
- [8] El algoritmo k-means aplicado a clasificación y procesamiento de imágenes: https://www.unioviado.es/compnum/laboratorios_py/kmeans/kmeans.html. (Fecha de acceso: 30/06/2021).