



CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

Detección de COVID-19 en radiografías de tórax

Nombre y Apellido del Alumno: Sebastián Iglesias

ÍNDICE

1. Estado de la cuestión	3
2. Definición del problema.....	3
3. Justificación del estudio.....	3
4. Alcances del trabajo y limitaciones.....	4
5. Hipótesis.....	4
6. Objetivos.....	5
7. Metodología a utilizar	5
A. PRE-PROCESAMIENTO de imágenes.....	5
B. Histograma de Gradientes Orientados.....	10
C. Clasificadores.....	13
I. Random Forest.....	13
II. SVM.....	15
8. Resultados	16
9. Conclusión	20
10. Trabajos Futuros.....	21
11. Referencias-Bibliográficas	22

1. ESTADO DE LA CUESTIÓN

El COVID-19 es una enfermedad infecciosa reconocida como pandemia por la Organización Mundial de la salud [1]. Una de las técnicas definidas por este ente para controlar la propagación del virus, es el seguimiento de contactos (“Contact Tracking”). Por lo tanto, un paso crítico para llevar a cabo esa medida es la detección efectiva y acertada de pacientes que contraigan la enfermedad. Tanto para recibir el tratamiento adecuado de manera pronta, como también aislarlos del público en general, para prevenir futuros contagios.

Una de las técnicas más popularizadas para la detección del COVID-19 es RT-PCR, reacción en cadena de la polimerasa con transcripción inversa, que analiza la producción de anticuerpos en respuesta a la infección ocasionada por dicho virus [2-3]. Este tipo de técnicas, propios de la serología tiene limitaciones, como por ejemplo la disponibilidad de equipos de testeo que provee dificultades a la hora de detectar el virus en un alto número de la población. Más allá de esto, el tiempo en que una de estas pruebas puede dar un resultado es entre algunas horas a hasta dos días. Inclusive, el resultado de estas pruebas, en la situación de emergencia mundial, es propensa a errores.

De esta manera, a partir de la necesidad de pruebas más veloces y con una menor propensión a errores, nacen estudios que involucran el análisis de imágenes de rayos x y tomografías utilizando visión artificial, específicamente en la región del pecho [4]. La mayoría de los pacientes con COVID-19 destacan opacidades recíprocas, multi-focales, similares a vidrio esmerilado, con una diseminación marginal en la etapa temprana y tardía de la infección [5]. El estilo de técnicas de visión artificial predilecto para estos casos, son las redes de aprendizaje profundo que pueden indicar características difíciles de singularizarse de la imagen original. El modelo predilecto para estos casos son las redes neuronales convolucionales [6].

DICOM es el formato estándar [9] en que se encuentran las radiografías “crudas” para realizar el análisis. A estas imágenes se les realiza un preprocesamiento, que suele incluir una primera etapa de remover partes innecesarias de la imagen (bordes y espacios oscuros al costado de los cuerpos), una segunda donde se le aplican técnicas de reducción de ruido y una tercera etapa donde se modifica el tamaño de éstas, respetando el “*aspect ratio*” de la imagen.

2. DEFINICIÓN DEL PROBLEMA

Este tipo de estudios utilizan normalmente redes neuronales convolucionales para la detección de COVID-19 [7-11]. Pero es poco el uso de otras herramientas de tratamiento de imágenes y detección de patrones en el nicho de radiografías de tórax en la detección de COVID-19. La intención de este trabajo es utilizar técnicas diferentes para realizar dicha detección.

3. JUSTIFICACIÓN DEL ESTUDIO

Este estudio tiene la intención de comparar dos clasificadores (“Random Forest” y SVM) que reciban como entrada el resultado de la técnica de Histograma de Gradientes Orientados para

radiografías de tórax, de pacientes con neumonía sin presencia del virus COVID-19, pacientes que si presentan el virus COVID-19 y pacientes sin ninguna de esas enfermedades. Sobre cada uno de los clasificadores entrenados que detecte si un individuo presenta o no la enfermedad, se calculan las siguientes métricas de evaluación: “accuracy”, “precisión”, “recall”, la curva ROC y AUC.

Un clasificador que permita la detección de una enfermedad, es una herramienta adicional al análisis médico hecho por un profesional en el área. Las pruebas PCR no siempre están disponibles, además de la velocidad de un resultado, junto al costo asociado de ese estilo de prueba impactan en la velocidad de tratamiento del paciente, junto a la logística de una clínica u hospital. La rápida respuesta y leve costo de un clasificador resulta beneficioso en este caso.

4. ALCANCES DEL TRABAJO Y LIMITACIONES

El conjunto de datos a utilizar en la investigación es provisto por “COVID-X”, una iniciativa argentina que utiliza modelos predictivos basados en imágenes radiográficas de tórax para la detección de COVID-19. El proyecto “COVID-X” provee a este trabajo de investigación con las imágenes anonimizadas.

Esta investigación no tiene la intención de realizar una aplicación ni pagina web que visualice el modelo, ni permita a un usuario ingresar imágenes para su posible detección. El resultado de este trabajo es una comparación en términos de la precisión en la detección de la enfermedad, partiendo de un mismo conjunto de datos basado en detección de características de imágenes.

5. HIPÓTESIS

Los modelos de Random Forest y SVM presentan las mismas métricas en términos de “accuracy”, “precision” y “recall” del modelo predictivo para la detección de COVID-19. Esas tres variables son métricas utilizadas para evaluar modelos de clasificación. “Accuracy” representa la cantidad de predicciones correctas del conjunto total de predicciones. “Precision” permite indicar que proporción de los resultados que fueron identificados como correctos son verdaderamente correctos. Y “recall” indica la proporción de los casos verdaderamente positivos fueron identificados como tales.

Siendo TP el número de predicciones positivas correctas para una clase, TN el número de predicciones negativas correctas para una clase, FP el número de predicciones que dieron positivas cuando el verdadero valor era negativo y FN el número de predicciones que dieron negativas cuando el verdadero valor era positivo

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

6. OBJETIVOS

Objetivos Generales:

- Obtener la efectividad del modelo predictivo asociado medido en términos de “accuracy”, “precision” y “recall”.

Objetivos específicos:

- Realizar un pre-procesamiento de las radiografías de tórax.
- Aplicar la técnica de Histogramas de Gradientes orientados.
- Entrenar y obtener métricas comparables para un modelo de “Random Forest” utilizando como entrada el resultado de la técnica de histogramas.
- Entrenar y obtener métricas comparables para un modelo de SVM utilizando como entrada el resultado de la técnica de histogramas.

7. METODOLOGÍA A UTILIZAR

En esta sección se describe las etapas del proceso para clasificar las imágenes de radiografías. Consta de un proceso manual de normalización, bajo la supervisión de un experto en el área, donde se recortan las imágenes dejando de lado áreas no relevantes al estudio. Seguido por un proceso automático de limpieza de las imágenes y ajuste de resolución para poder obtener sus características a partir del método de histogramas de gradientes orientados. Finalmente esas características son procesadas por un clasificador.

A. PRE-PROCESAMIENTO DE IMÁGENES

Las imágenes obtenidas de radiografías de tórax provienen de diversos hospitales, realizadas con máquinas diferentes en pacientes en múltiples posiciones, deben ser normalizadas para su análisis automático. La normalización comienza por realizar un recorte manual de las imágenes, con un profesional en el área presente, para que las imágenes contengan, en lo posible únicamente el área de la imagen ocupada por los pulmones.

La Fig. 1 muestra un ejemplo de una radiografía frontal del tórax original. En la Fig. 1 aparecen los bordes negros que dificultan la interpretación automática, la Fig. 2 muestra la misma imagen recortada con la ayuda de un experto.



Fig. 1 Radiografía frontal de tórax que muestra bordes negros junto a notaciones utilizadas por expertos en el área para determinar que lado corresponde al izquierdo o derecho de paciente.

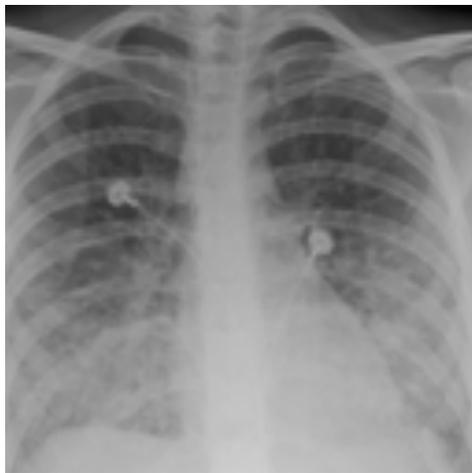


Fig. 2 Misma imagen de la *Fig 1* recortada dejando de lado áreas que no involucren a los pulmones

Los bordes oscuros o las secciones de la imagen que no aportan información pertinente para este estudio se eliminan, y dan lugar a poder trabajar mejor sobre el contraste de las imágenes al convertirlas en una escala de gris. Para eliminar el ruido y poder percibir con mayor facilidad las opacidades producidas por la enfermedad de COVID-19 se debe aplicar la técnica de filtro bilateral.

Esta técnica permite “suavizar” las imágenes preservando bordes a través de una combinación no lineal de valores en un área de la imagen. Combina los valores de color basados en la cercanía geométrica y la similitud fotométrica, priorizando valores cercanos antes que lejanos teniendo en cuenta su rango y dominio. La *Fig. 3* ejemplifica el uso de este filtro comparando dos imágenes, la izquierda sin aplicarse el filtro, y la imagen derecha aplicándole el filtro bilateral.

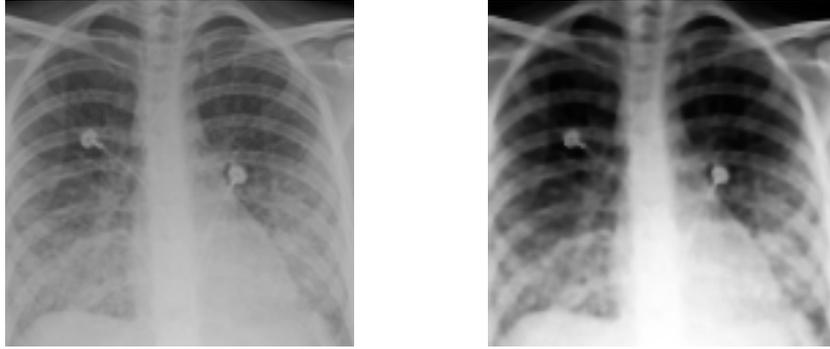


Fig. 3 Imagen de la izquierda muestra una radiografía de tórax sin aplicar ningún filtro. Imagen de la derecha muestra una radiografía habiendo aplicado el filtro.

Esta técnica de filtrado es similar al concepto de filtro Gaussiano. Este filtro es un promedio ponderado de la intensidad de posiciones adyacentes con un “peso” descendiente según la distancia del punto con respecto a la posición centro. Es necesario mencionar en primera instancia el filtro de Gauss:

$$GB[I]_p = \sum_{q \in S} G_{\sigma}(|p - q|) I_q$$

La fórmula indica a **p** y **q** como posiciones de los píxeles, **I** siendo la imagen y por último $G_{\sigma}(x, y)$ es una función gaussiana de dos dimensiones.

$$G_{\sigma}(x, y) = \frac{1}{2 \pi \sigma^2} * e^{-\frac{x^2 + y^2}{2 \sigma^2}}$$

La función G_{σ} decrementa la influencia de los píxeles más lejanos, definidos por una distancia $G_{\sigma}(|p - q|)$, y σ funciona como un parámetro para definir el tamaño del espacio de coordenadas.

Por lo tanto, con un filtro gaussiano, solamente se consideran píxeles “ceranos” al filtrar, sin incluir la intensidad de los mismos o si forman parte de un borde. Esto suele afectar negativamente al análisis de imágenes, porque al difuminar los bordes se pierde el sentido de la imagen, en este caso particular, no quedaría claro donde comienza un pulmón en la radiografía. Entonces, este tipo de filtro deja de lado elementos cruciales para el análisis. La *Fig. 4* muestra el uso únicamente del filtro gaussiano, con notables diferencias al filtro bilateral. El filtro gaussiano aplicado en la imagen derecha, muestra bordes de menor nitidez que el filtro bilateral.

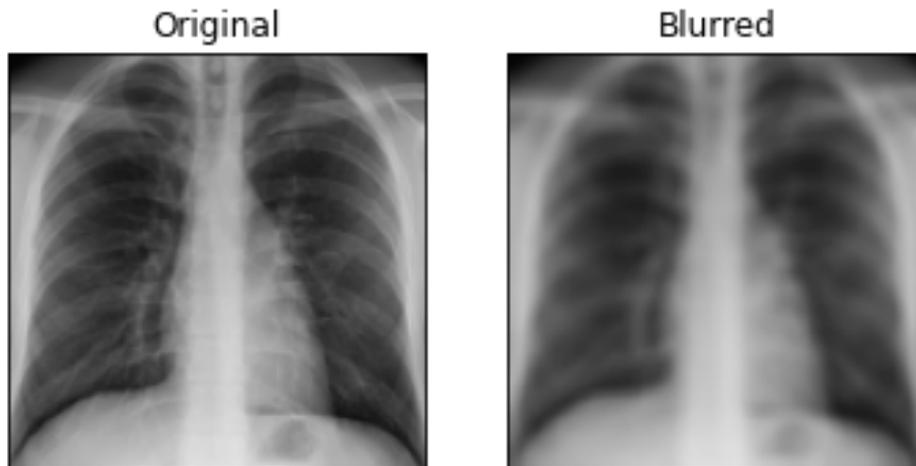


Fig. 4 Imagen de la izquierda muestra una radiografía de tórax sin procesar e imagen de la derecha muestra la misma imagen habiendo aplicado el filtro de Gauss produciendo un efecto de borronado

El filtro bilateral, considera un filtro gaussiano extra que introduce el concepto de diferencia entre intensidad de píxeles. Este filtro tiene una función gaussiana relativa al espacio donde solamente los píxeles cercanos difuminan, y una segunda función gaussiana que se concentra en difuminar únicamente píxeles de intensidades similares al pixel ubicado en el centro. Esto permite preservar los bordes, ya que los píxeles que formen parte de un borde van a presentar una gran variación de intensidad.

$$BF[I]_p = \frac{1}{W_p} \sum_{q \in S} G_{\sigma_s} (|p - q|) G_{\sigma_r} (|I_p - I_q|) I_q$$

BF es la función de filtro bilateral con W_p siendo un factor de normalización:

$$W_p = \sum_{q \in S} G_{\sigma_s} (|p - q|) G_{\sigma_r} (|I_p - I_q|)$$

A comparación del filtro gaussiano, se agregan dos términos:

$$G_{\sigma_r} (I_p - I_q) I_q$$

$$\frac{1}{W_p}$$

G_{σ_s} continúa representando lo mismo que en la función de filtro gaussiano, es una función gaussiana espacial que disminuye la influencia de los píxeles más distantes. G_{σ_r} es el término agregado que disminuye la influencia de píxeles I_q que tengan un rango fotométrico diferente de I_p .

Los parámetros del filtro bilateral son σ_s y σ_r , siendo el parámetro de espacio y de rango respectivamente. A su vez, los parámetros que posee la implementación de “OpenCV” son:

- **SigmaColor:** una variable de tipo entero que representa el filtro sigma en el espacio de color. Un valor alto implica que una mayor cantidad de “colores” en el vecindario de píxeles va a ser “mezclado” al difuminar la imagen, dando por resultado áreas de la imagen con color similar.
- **SigmaSpace:** una variable de tipo entero representando el filtro sigma en el espacio de coordenadas. Un valor alto implica que píxeles más distantes tienen posibilidad de influir en el resultado siempre y cuando presenten una diferencia chica de color e intensidad.
- **D:** Diámetro de cada “vecindad” de un píxel al momento de aplicar el filtro.

El algoritmo subyacente de esta implementación consta de los siguientes pasos:

- a. Para cada píxel de la imagen, se calcula el promedio ponderado de sus “vecinos”.
- b. Cada “vecino” se lo pondera por el componente espacial que va penalizando píxeles más distantes y también por el componente de rango que penaliza píxeles con diferente nivel de intensidad. Esta combinación permite asegurar que solo píxeles cercanos similares contribuyan al resultado final.

Esto permite que bordes muy oscuros o claros, como son comúnmente vistos en radiografías de tórax alteren de manera negativa el contraste resultante. La *Fig. 5* muestra los efectos de diferentes valores de parámetros para el filtro bilateral, pudiendo obtener mayor o menor nitidez de la imagen aplicando el suavizado del filtro.



Fig. 5 Una misma radiografía de tórax aplicando filtro bilateral con diferentes valores de sigmas. De izquierda a derecha: (a) D: 3 SigmaColor: 25 SigmaSpace: 25 (b) D 4 SigmaColor: 50 SigmaSpace: 50 (c) D:5 SigmaColor: 75 SigmaSpace:75

Por último se debe modificar el tamaño de la imagen a una resolución estándar respetando el ratio de la misma. Se deben convertir las imágenes a una resolución de 128x128 para ser usadas como entrada en la técnica de histogramas de gradientes orientados.

B. HISTOGRAMA DE GRADIENTES ORIENTADOS

El histograma de gradientes orientados o HOG es un descriptor de características. Un descriptor de características es una representación de una imagen o una parte de esta, que la simplifica al extraer de la imagen información que considera relevante, dejando de lado información que no considera importante. La *Fig 6* muestra el resultado de procesar una imagen de radiografía de tórax por medio del método de histogramas de gradientes orientados.

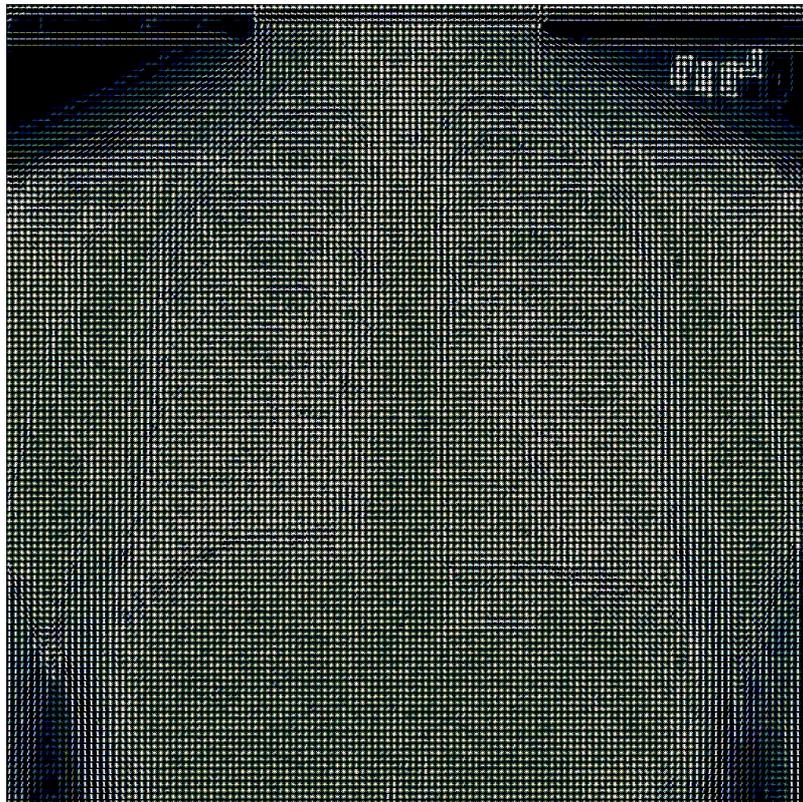


Fig. 6 Visualización de una radiografía de tórax procesada por medio del histograma de gradientes orientados

HOG utiliza la distribución o histograma de las direcciones de gradientes (o también llamados gradientes orientados) como características, de ahí su nombre. Los gradientes se conforman por las derivadas de los ejes x e y, estos suelen ser útiles principalmente para la identificación de bordes o regiones con cambios abruptos de intensidades de colores.

La manera de identificar la presencia de la enfermedad COVID-19 en radiografías frontales de toráx es a partir de regiones opacas similares a vidrio esmerilado que resaltan como cambios abruptos del típico color oscuro que presentan los pulmones.

Para calcular la magnitud y dirección de los gradientes horizontales y verticales, se utilizan las siguientes formulas para cada uno de los pixeles:

$$G_x(r, c) = I(r, c + 1) - I(r, c - 1)$$

$$G_y(r, c) = I(r - 1, c) - I(r + 1, c)$$

Donde r y c representan las filas y columnas respectivamente de la imagen I . Obteniendo el gradiente es necesario calcular la magnitud y ángulo de cada pixel:

$$\text{Magnitud}(\mu) = \sqrt{G_x^2 + G_y^2}$$

$$\text{ángulo}(\theta) = \left| \tan^{-1} \left(\frac{G_y}{G_x} \right) \right|$$

Se obtiene como resultado una dos matrices, una de magnitud y otra de ángulos dividida en grupos de celdas de $n \times n$ que conforman un bloque. Dentro de cada bloque, se calcula un histograma con m secciones, cada sección contiene un rango de ángulos 20 grados.

$$\text{Número de secciones} = 9 \text{ (desde } 0^\circ \text{ a } 180^\circ)$$

$$\text{Tamaño de sección } (\Delta\theta) = \frac{180^\circ}{\text{Número de secciones}}$$

Cada una de las secciones tendrá como límites:

$$[\Delta\theta \cdot j, \Delta\theta \cdot (j + 1)]$$

$$\text{Centro de una sección } j = C_j = \Delta\theta (j + 0.5)$$

Valor									
Sección	0	20	40	60	80	100	120	140	160

Fig. 7 Representación de un histograma con 9 secciones.

Para cada celda en el bloque, se debe calcular el valor de j y $j + 1$ utilizando las fórmulas:

$$j = \left\lfloor \left(\frac{\theta}{\Delta\theta} - \frac{1}{2} \right) \right\rfloor$$

$$V_j = \mu \cdot \left(\frac{\theta}{\Delta\theta} - \frac{1}{2} \right)$$

$$V_{j+1} = \mu \cdot \left(\frac{\theta - C_j}{\Delta\theta} \right)$$

Se utiliza un arreglo para representar una sección para un bloque, los valores V_j y V_{j+1} se colocan en el arreglo siguiendo el índice j y $j + 1$ calculado para cada pixel. Al haber

computado el histograma para cada bloque, se combinan los bloques entre sí para obtener bloques de menor tamaño, f_{bi} . La combinación se realiza seleccionando regiones de pixeles que se solapen entre bloques:

$$f_{bi} = [b_1, b_2, b_3, \dots, b_n]$$

Estos bloques de menor tamaño se normalizan, siendo ε un número significativamente pequeño para evitar una división por 0.

$$f_{bi} \leftarrow \frac{f_{bi}}{\sqrt{\|f_{bi}\|^2 + \varepsilon}}$$

$$k = \sqrt{b_1^2 + b_2^2 + b_3^2 + \dots + b_n^2}$$

$$f_{bi} = \left[\left(\frac{b_1}{k}\right), \left(\frac{b_2}{k}\right), \left(\frac{b_3}{k}\right), \dots, \left(\frac{b_n}{k}\right) \right]$$

Esta normalización se realiza para reducir el efecto de cambios de contraste entre imágenes similares. La implementación utilizada en esta investigación es la librería “hog” de “scikit-image”. Los parámetros que utiliza son:

- **Image:** Una imagen representada por un arreglo sobre la cual se aplica el algoritmo de histograma de gradientes orientados.
- **orientations:** Número de secciones de orientaciones.
- **pixels_per_cell:** Tamaño en pixeles de una celda.
- **cells_per_block:** Número de celdas en cada bloque.

La Fig. 8 representa visualmente las características obtenidas por medio de histogramas de gradientes orientados para las imágenes modificadas a una resolución de 128x128. Tener un mayor o menor número de pixeles y celdas por bloque afecta el número de gradientes obtenidos. Mientras que el número de orientaciones afecta la forma final de cada una de las celdas dentro de los bloques.

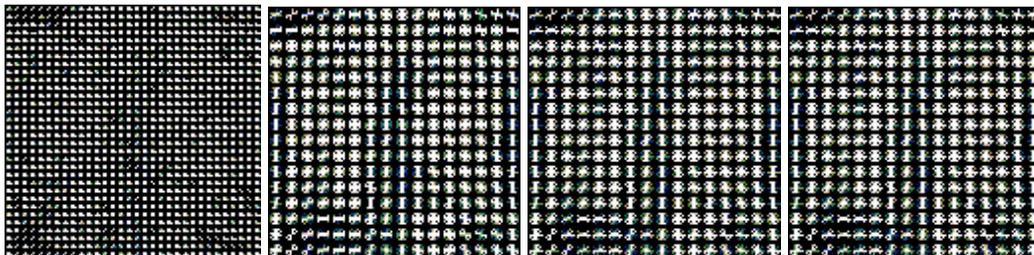


Fig. 8 Variaciones de diferentes parámetros sobre una radiografía de tórax con una resolución de 128 píxeles x 128 píxeles

C. CLASIFICADORES

En esta sección se describen los métodos de clasificación utilizados, “Random Forest” y SVM. Ambos clasificadores reciben como entrada las mismas características resultantes del método de histogramas de gradientes orientados junta a un conjunto de parámetros propia de cada método adecuada para el análisis en cuestión.

I. *RANDOM FOREST*

El clasificador de Random Forest consiste de un gran número de árboles de decisión individuales que operan como un “ensemble” o conjunto. Cada árbol individual, devuelve un resultado para una clasificación en particular, y la clase obtenida en la mayoría de los árboles es la predicción final del clasificador.

El concepto fundamental detrás del funcionamiento de este clasificador, es la “sabiduría de las masas”. Es un término utilizado en las ciencias de datos que representa a un gran número de modelos relativamente no correlacionados, en este caso particular arboles de decisión, operando como un comité, superará a cualquiera de los modelos individuales. Siendo un factor de suma importancia la no correlación entre modelos.

Existe un concepto similar en el área de finanzas e inversiones, donde un individuo o compañía buscan generar portfolios compuestos por elementos y herramientas financieras con baja correlación entre si. La baja de alguna de estas herramientas es compensada por otras que estén en alza.

El mismo concepto se aplica con el error de un único árbol de decisión. Siempre y cuando los arboles sean lo suficientemente no correlacionados y no todos tengan un error en la misma dirección, el error de ciertos modelos se compensara con el de muchos otros. El clasificador Random Forest utiliza dos técnicas para evitar arboles de decisión correlacionados entre si, Bagging y Aleatoriedad de características.

En el primer lugar, la técnica de Bagging se basa en que los arboles de decisión, son propensos a ser altamente sensitivos a los datos con los que se entrenan. Pequeños cambios en el conjunto de entrenamiento de un árbol resultan en estructuras significativamente diferentes.

La técnica de Bagging consiste en tomar ventaja de esto, y permitirle a cada árbol individual del conjunto tomar una muestra aleatoria del conjunto de datos de entrenamiento, resultando en árboles diferentes. La toma de la muestra, es con reemplazo, implica que no se separan sub-conjuntos específicos para cada árbol. Sea una muestra de tamaño N , cada árbol recibe de igual forma un conjunto de entrenamiento de tamaño N pero con diferentes valores.

En segundo lugar, la aleatoriedad de características, es una técnica que modifica la lógica utilizada al momento de realizar un Split en un nodo del árbol. Normalmente, un nodo del árbol considera todas las posibles características de un conjunto de datos, y realiza el Split teniendo en cuenta la característica que genera una mayor separación entre los nuevos nodos, comúnmente calculado con el índice gini o de entropía.

Esta técnica limita la cantidad de características que cada árbol puede utilizar para realizar la separación del nodo. El conjunto de características del que pueden seleccionar es también azaroso y esto fuerza una mayor variación entre los árboles del conjunto de modelos, diversificando los resultados y resultando en una menor correlación.

La librería de sklearn de RandomForestClassifier ofrece un conjunto de parámetros para modificar el clasificador:

- **N_estimators:** Este es el número de estimadores (árboles de decisión) que pertenecerán al conjunto de Random Forest. Con un alto número de estimadores, se incrementa el nivel de “generalización” del clasificador.
- **Criterion:** Al momento de separar un nodo en 2 o más nodos, se debe utilizar una característica que genere la mayor separación entre los conjuntos de los futuros nodos. El criterio bajo el cual se realiza esa separación puede ser usando el indicador de impureza de Gini o la entropía.
- **Max_depth:** Es el nivel máximo de profundidad que puede tener un árbol de decisión. Normalmente, cuantos más niveles de profundidad tenga un árbol, es más propenso a producir overfitting. Al ser un clasificador ensemble, el error generado por una gran profundidad es compensado por otros árboles.
- **Max_features:** El número de características a considerar cuando se busca el mejor criterio de Split de un nodo. Esto permite la aleatoriedad de características en diferentes árboles y disminuye la correlación entre los mismos.
- **Min_samples_leaf:** Es el número mínimo de muestras requeridas en un nodo para considerarse un nodo hoja. Permite evitar overfitting, cuando un Split de un nodo no se realiza por no cumplir la condición de tener la mínima cantidad de muestras, se dice que realiza un “alisado”. Queda como resultado el “promedio” de los posibles nodos hijos que no se crearon.
- **Min_samples_split:** El número mínimo de muestras requeridas para realizar un Split de un nodo.
- **Bootstrap:** parámetro booleano para realizar o no el Bagging. De ser falso, se utiliza todo el conjunto de datos de entrenamiento para cada árbol, de ser verdadero se realiza el Bagging

Para determinar que valores de parámetros son convenientes, o con cuales se obtiene el mejor modelo, se hace uso de la técnica de grid search. Consiste en realizar una combinatoria de

múltiples valores de parámetros, entrenar un modelo por cada combinación de parámetros y a partir de los resultados de cada modelo contra un conjunto de prueba, se selecciona al “mejor modelo” con sus parámetros.

II. SVM

Support vector machine o SVM, es un modelo lineal para resolver problemas de clasificación y/o regresión. La base del algoritmo es crear una línea o en su defecto un hiperplano que separe los datos presentados en un plano n-dimensional, en diferentes clases.

Los hiperplanos se consideran también como límites de decisión, que clasifican puntos que representan datos en una determinada clase en un espacio multidimensional. Puntos que queden de un cierto lado del hiperplano se atribuyen a diferentes clases. En dos dimensiones, un hiperplano sería una línea, en tres un plano y en más dimensiones un hiperplano.

Se utiliza la implementación de “sklearn” para SVM que tiene los siguientes parámetros:

- **C:** El parámetro C le indica la optimización de SVM que tanto se desea evadir fallar en clasificar cada elemento del conjunto de prueba. Para valores de C elevados, la optimización obtenida tendrá un hiperplano de menor margen siempre y cuando el hiperplano clasifique correctamente los puntos de entrenamiento. Para valores bajos de C, se obtiene un hiperplano de mayor margen, aun cuando el hiperplano falle en clasificar ciertos puntos. El margen es calculado por medio del error de la distancia al cuadrado de los puntos al hiperplano.
- **Gamma:** Define que tan lejana es la influencia de un único punto de entrenamiento. Para valores bajos, significa que cada punto tiene un alcance amplio e inversamente, para valores altos de gamma cada punto tiene un alcance corto de influir.
- **Kernel:** La función que utiliza SVM se define como kernel. Los kernel reciben una cierta información como entrada y la transforman en otra forma requerida. Puede ser lineal, polinómica, radial o sigmoidea.

8. RESULTADOS

En esta sección se muestran los resultados obtenidos de ambos clasificadores.

La *Fig. 9* muestra la tabla resultante del mejor clasificador de Random Forest. Para este análisis se utilizó la librería de sklearn RandomForestClassifier como clasificador en conjunto a la librería GridSearchCV. Esta última permite la búsqueda de la mejor combinación de parámetros de filtro bilateral, histograma de gradientes orientados y “Random Forest” por medio de una búsqueda “Grid” haciendo uso de la técnica ‘cross-validation’.

Random forest #1	Sin Enfermedad	Neumonia	COVID-19
Accuracy	0.76	0.70	0.73
Recall	0.82	0.65	0.72
Precision	0.76	0.73	0.81

Fig. 9 Tabla que muestra los mejores resultados obtenidos en términos de Accuracy, Recall y Precision para el modelo de “Random Forest” al momento de clasificar las radiografías en las clases Normal (paciente que no presenta ni neumonía ni COVID-19), Neumonía y COVID-19

La *Fig.10* muestra la curva ROC y el AUC de “Random Forest”, indicada como “ROC curve (área)” para cada una de las clasificaciones, sin enfermedad, con neumonía y con COVID-19. La línea azul representa la clasificación sin enfermedad, la amarilla de neumonía y la verde COVID-19. Este modelo clasifica correctamente con mayor precisión a pacientes sin la enfermedad que con neumonía o COVID-19, pero de todas formas presenta valores aceptables para las otras dos clasificaciones.

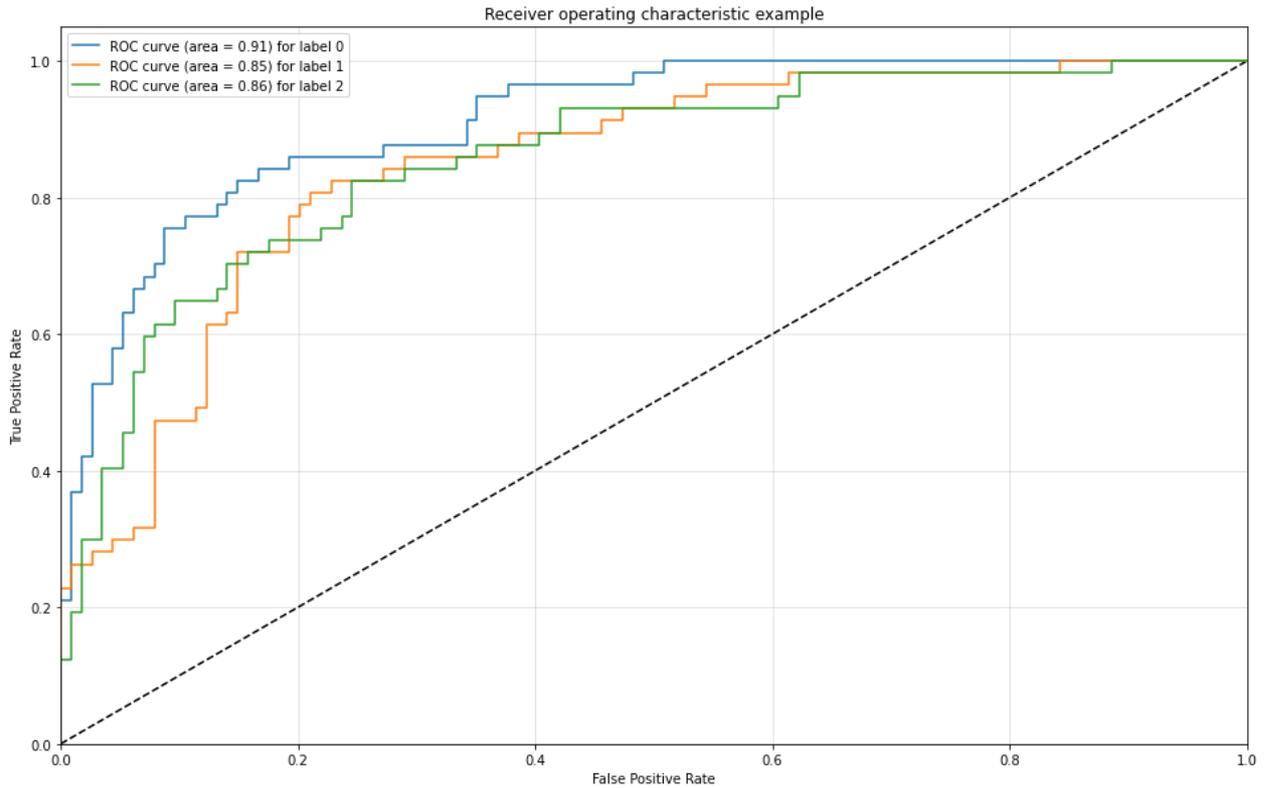


Fig. 10 Curva de ROC del mejor clasificador obtenido luego del proceso de “GridSearch” para el clasificador “RandomForest”

La Fig. 11 muestra la matriz de confusión para el mejor clasificador obtenido de “Random Forest”. Para obtener esta matriz se utilizó un conjunto de datos con igual número de cada clase, 57 para cada clase, 171 datos en total.

		Clases obtenidas como resultado del modelo		
		Sin Enfermedad	Neumonía	COVID-19
Clases real	Sin enfermedad	48	7	2
	Neumonía	7	39	11
	COVID-19	11	14	32

Fig. 11 Matriz de confusión para el mejor modelo obtenido con el clasificador de random forest.

La Fig. 12 muestra la tabla resultante del mejor clasificador de SVM. Para este análisis se utilizó la librería de sklearn SVM como clasificador en conjunto a la librería GridSearchCV. Esta última permite la búsqueda de la mejor combinación de parámetros de filtro bilateral, histograma de gradientes orientados y “SVM” por medio de una búsqueda “Grid” haciendo uso de la técnica ‘cross-validation’.

SVM	Sin enfermedad	Neumonia	COVID-19
Accuracy	0.83	0.78	0.83
Recall	0.86	0.80	0.77
Precision	0.76	0.78	0.74

Fig. 12 Tabla que muestra los mejores resultados obtenidos en términos de Accuracy, Recall y Precision para el modelo de SVM al momento de clasificar las radiografías en las clases Normal (paciente que no presenta ni neumonía ni COVID-19), Neumonía y COVID-19

La *Fig.13* muestra la curva ROC y el AUC para cada una de las clasificaciones, sin enfermedad, con neumonía y con COVID-19. La línea azul representa la clasificación sin enfermedad, la amarilla de neumonía y la verde COVID-19. Este modelo clasifica correctamente con un nivel de precisión similar a pacientes sin la enfermedad y con COVID-19 a diferencia de pacientes con neumonía.

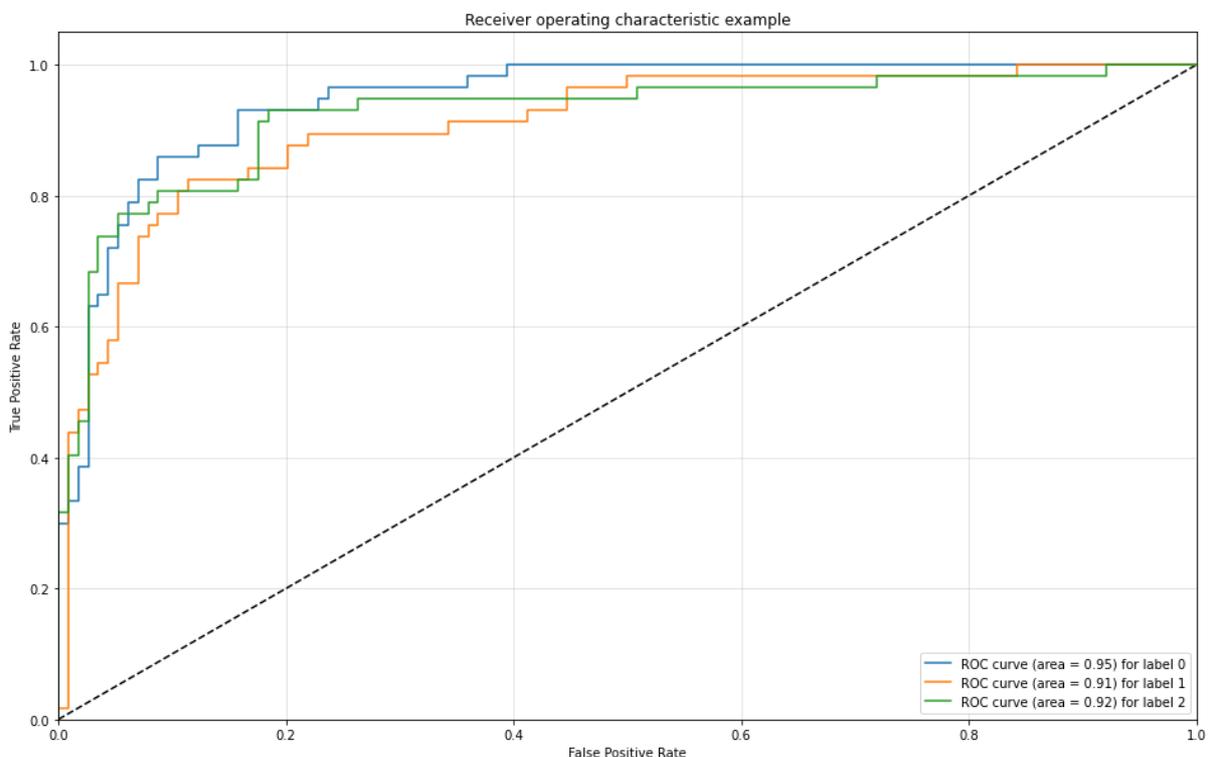


Fig. 13 Curva de ROC del mejor clasificador obtenido luego del proceso de GridSearch para el clasificador SVM

La *Fig. 14* muestra la matriz de confusión para el mejor clasificador obtenido de "SVM". Para obtener esta matriz se utilizó un conjunto de datos con igual número de cada clase, 57 para cada clase, 171 datos en total.

		Clase obtenida como resultado del modelo		
		Sin enfermedad	Neumonía	COVID-19
Clase real	Sin enfermedad	43	7	7
	Neumonía	8	36	13
	COVID-19	10	9	38

Fig. 14 Matriz de confusión para el mejor modelo obtenido con el clasificador de SVM.

9. CONCLUSIÓN

En el área de detección de enfermedades a partir de imágenes médicas, las redes neuronales son la herramienta normalmente utilizada, en este trabajo, se utilizan otros métodos de clasificación alternativos, que son “Random Forest” y SVM.

Analizando los resultados, se concluye que métodos “tradicionales” como son las “Support Vector Machines” o “Random Forests”, pueden obtener también buenos resultados. A su vez, el extractor de características de histogramas de gradientes orientados, comúnmente utilizado en el área de detección de objetos en imágenes, probó ser útil al aplicarse en imágenes médicas.

Ambos modelos resultantes recibieron una misma entrada de datos, dando resultados similares, con SVM siendo levemente más preciso al detectar la presencia de las lesiones ocasionadas en los pulmones por COVID-19 que “Random Forest”.

Un clasificador que permita la detección de una enfermedad, en este caso COVID-19, es una herramienta más que puede sumarse al análisis médico hecho por un profesional en el área. Tiene el beneficio de ofrecer un costo bajo en comparación al costo y disponibilidad de pruebas PCR junto a la velocidad de su resultado. El impacto que puede tener esta herramienta se incrementa al tener en cuenta el costo logístico de una clínica u hospital, dado que un caso positivo de COVID-19 requiere de un aislamiento. Y el costo de no lograr detectar a un paciente como positivo cuando verdaderamente lo es, es elevado al acarrear posibles contagios y mayores probabilidades de internación de nuevos pacientes.

10. TRABAJOS FUTUROS

Existen varios puntos, mayormente en el pre-procesamiento de los datos, que pueden considerarse como trabajos futuros.

- La detección de los pulmones dentro de las radiografías de manera automática.

El detectar y reconocer el área dentro de una radiografía requiere de la presencia de un profesional en el área, especialmente para radiografías de tórax que presentan un nivel de ruido considerable. El realizar esta detección de manera automática, ahorraría mucho tiempo al momento de procesar las imágenes y realizar los cortes apropiados en cada una, dejando de lado cualquier otra área de la radiografía que no contenga pulmón

- Incorporar información extra además de las características obtenidas del histograma de gradientes orientados

Considerar incluir en el modelo predictivo diversos datos, pudiendo provenir de la misma imagen de tórax o información adicional específica del paciente. Se podría incluir como entrada del modelo la imagen pasado por diferentes filtros o múltiples estadísticas de la misma (analizando varianza o promedio de diferentes valores de color en cada pixel de la imagen) como también agregar información médica pertinente como nivel de oxígeno en sangre, sexo y edad del paciente del cual se obtuvo la radiografía.

11. REFERENCIAS-BIBLIOGRÁFICAS

- [1] WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. (n.d.). Retrieved July 29, 2020, from <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>
- [2] Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., & Tan, W. (2020). Detection of SARS-CoV-2 in Different Types of Clinical Specimens. In *JAMA - Journal of the American Medical Association* (Vol. 323, Issue 18, pp. 1843–1844). American Medical Association. <https://doi.org/10.1001/jama.2020.3786>
- [3] Emery, S. L., Erdman, D. D., Bowen, M. D., Newton, B. R., Winchell, J. M., Meyer, R. F., Tong, S., Cook, B. T., Holloway, B. P., McCaustland, K. A., Rota, P. A., Bankamp, B., Lowe, L. E., Ksiazek, T. G., Bellini, W. J., & Anderson, L. J. (2004). Real-Time Reverse Transcription-Polymerase Chain Reaction Assay for SARS-associated Coronavirus. *Emerging Infectious Diseases*, 10(2), 311–316. <https://doi.org/10.3201/eid1002.030759>
- [4] Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Chen, Y., Su, J., Lang, G., Li, Y., Zhao, H., Xu, K., Ruan, L., & Wu, W. (2020). Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. *Applied Intelligence*, 2019, 1–5. <http://arxiv.org/abs/2002.09334>
- [5] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., ... Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223), 497–506. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)
- [6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. <http://code.google.com/p/cuda-convnet/>
- [7] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., Xu, B., MedRxiv, M. C., 2020, U., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B. (2020). A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). *MedRxiv*, 1–19. <https://doi.org/10.1101/2020.02.14.20023028>
- [8] Duchesne, S., Gourdeau, D., Archambault, P., Chartrand-Lefebvre, C., Dieumegarde, L., Forghani, R., Gagne, C., Hains, A., Hornstein, D., Le, H., Lemieux, S., Levesque, M.-H., Martin, D., Rosenbloom, L., Tang, A., Vecchio, F., & Duchesne, N. (2020). Tracking and Predicting Covid-19 Radiological Trajectory Using Deep Learning on Chest X-Rays: Initial Accuracy Testing. *MedRxiv*, 1(418), 2020.05.01.20086207. <https://doi.org/10.1101/2020.05.01.20086207>
- [9] Ramadhan, M. M., Faza, A., Lubis, L. E., Yunus, R. E., Salamah, T., Handayani, D., Lestariningsih, I., Resa, A., Alam, C. R., Prajitno, P., Pawiro, S. A., Sidipratomo, P., & Soejoko, D. S. (2020). *Fast and accurate detection of Covid-19-related pneumonia from chest X-ray images with novel deep learning model*. <http://arxiv.org/abs/2005.04562>
- [10] Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). *Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network*. <http://arxiv.org/abs/2003.13815>
- [11] Asif, S., Wenhui, Y., Jin, H., Tao, Y., & Jinhai, S. (2020). *Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Networks*.