

Parametrización De Un Modelo De Red Neuronal Convolutiva Para La Clasificación De Cáncer En Ecografías De Tiroides Según La Escala EU-TIRADS.

Arias T. Yhary Estefania, Ortiz B. Ariel
yarias46@unab.edu.co, aortiz875@unab.edu.co
Universidad Autónoma de Bucaramanga

Resumen—La detección de cáncer de tiroides es un proceso que en la actualidad se realiza mediante la interpretación manual que realizan radiólogos especialistas, estas se clasifican utilizando una prueba de tamizaje (discriminatoria) conocida como EU-TIRADS 2017 [2], que determina el grado de malignidad del nódulo tiroideo. La escasez de profesionales y la creciente demanda de este tipo de estudios plantea el problema de la automatización a través de algoritmos de aprendizaje de máquina como los basados en Deep Learning y específicamente, las Redes Neuronales Convolutivas, que han sido probadas anteriormente con éxito para la clasificación de otro tipo de imágenes médicas.

En un trabajo anterior, con un dataset de 2000 imágenes balanceado entre 4 categorías (TI-RADS2 - TI-RADS5) se logró una medida de precisión (accuracy) cercana del 65% y una pérdida logarítmica (cross-entropy loss) cercana a 0.78. Sin embargo, este artículo plantea el estudio exploratorio para una posible optimización del algoritmo a través de diferentes pruebas medibles en su parametrización.

Las variables que serán ajustadas son: El número de capas convolutivas, el tamaño de la máscara de convolución, las funciones de activación, el número de neuronas en la capa densa, el uso de más capas densas para el aprendizaje, el uso de dropouts aleatorios para controlar el sobreajuste (overfitting), entre otros. La medición comparativa se realiza a través de los valores de precisión, pérdida, la matriz de confusión, y el área bajo la curva ROC. Al final del documento se describe la mejor combinación de los parámetros evaluados y las observaciones pertinentes.

I. INTRODUCCIÓN

Actualmente en su mayoría, las lecturas y procedimientos médicos se realizan de manera manual, procesos como identificación de nódulos tiroideos, por medio de una ecografía, este tipo de procedimientos presentan una alta demanda en la verificación y diagnóstico de cada paciente. Según estudios realizados por el Instituto Nacional de Cancerología [1] los casos más comunes en cuanto a las alteraciones en la glándula tiroidea, en su mayoría están presentes en mujeres, dado que en el país por cada ocho mujeres identificadas con problemas en la tiroides un hombre presenta el mismo diagnóstico. Las principales localizaciones de incidencias de cáncer entre los hombres fueron próstata, estómago, pulmón, colon-recto y ano, y linfomas no Hodgkin.

En mujeres, las principales localizaciones fueron mama, cuello del útero, colon-recto y ano, estómago y tiroides [1].

Este es uno de los procedimientos más realizados en Colombia debido que se presentan 2.201 casos de Cáncer en la Tiroides por cada 100.000 habitantes.

La figura 1 resume los últimos datos publicados de las 10 principales localizaciones de incidencia estimada anual y tasas ajustadas por edad en Colombia entre el 2007 y el 2011 [1].

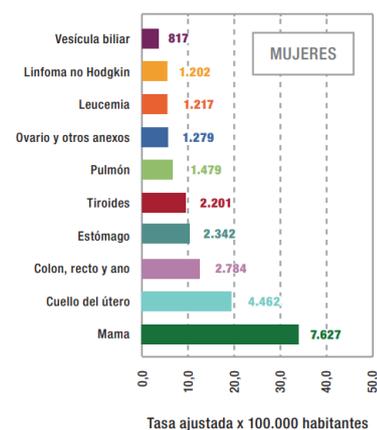


Figura 1-Incidencia de cáncer estimada anual en números absolutos y tasas ajustadas por edad, 10 principales localizaciones, hombres y mujeres, Colombia 2007-2011 [1]

El examen para identificar problemas en la tiroides se realiza por medio de un ultrasonido, este examen utiliza ondas sonoras para producir imágenes de la glándula Tiroidea en el cuello. No usa radiación ionizante y se utiliza generalmente para evaluar bultos o nódulos que se refiere a cualquier crecimiento anormal de las células tiroideas que forman un tumor dentro de esta glándula y estos son detectados durante un examen físico u otros exámenes por imágenes.

El objetivo de una evaluación ecográfica es distinguir los nódulos benignos que pueden pasar a observación o malignos que requieren de un tratamiento adicional.

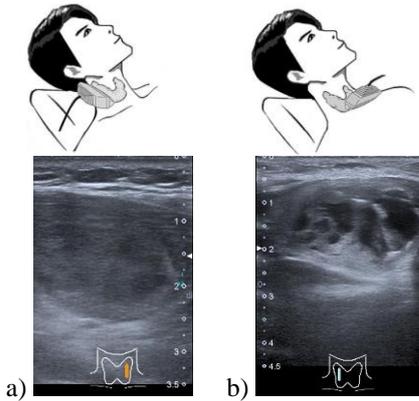


Figura 2- Procedimiento de adquisición. a) Adquisición de plano longitudinal de lóbulo derecho de la tiroides. (b) Adquisición del plano transversal del lóbulo tiroideo izquierdo.

2. PROBLEMA

En el proceso de lectura de imágenes en ultrasonidos se aplica una prueba de tamizaje que se encarga de clasificar el riesgo carcinoma de Tiroides llamada TIRADS [1] (Thyroid Imaging Reporting and Data System - Sistema de datos e informes de imágenes de la Tiroides) que consiste en un sistema de clasificación ecográfica que describe los hallazgos encontrados en un Nódulo Tiroideo. Este sistema fue propuesto por Horvath et al en 2009. Este TIRADS se modifica según ciertos criterios, en este caso EU-TIRADS [1] es una de las pruebas más flexibles diseñadas en Europa.

Esta prueba de tamizaje es de fácil implementación para especialistas debido a la escala que se divide de 1 a 5, proporcionando criterios de selección, siendo TIRADS 1 (normal, 0% riesgo de cáncer), TIRADS 2 (benigno, ≥ 0 riesgo de cáncer, quiste puro) TIRADS 3 (2% a 4% riesgo de cáncer), TIRADS 4 (6% a 17% riesgo de cáncer), TIRADS 5 (26% a 87% maligno, alto riesgo de cáncer) [1]. Este sistema de clasificación se realiza manualmente lo que ocasiona demoras a para generar un diagnóstico correcto basado en las métricas para cada paciente, por ello es necesario implementar nuevas tecnologías que brinden apoyo a los procesos de lectura en imágenes e identificar de manera temprana, anomalías o cáncer en la tiroides con el propósito generarle a cada paciente un tratamiento de acuerdo con el problema que presenta y evitar muertes por cáncer en la tiroides.

3. PROCEDIMIENTO PROPUESTO

Se desarrolló un estudio de concordancia observacional con el propósito de implementar un sistema de inteligencia artificial para la clasificación de ecografías de nódulos tiroideos en uso de una prueba de tamizaje que se aplican en diferentes glándulas para identificar el riesgo de cáncer en el paciente, dependiendo de la zona que se evalúan su nombre varía, para la tiroides esta prueba de tamizaje recibe el nombre de TI-RADS

a la cual se le han realizado modificaciones para generar un diagnóstico con mayor precisión.

En la construcción del modelo de procesamiento y clasificación se da usos a la escala europea EU-TIRADS 2017 que fue su última versión, para ello se diseña una base de datos con más de mil imágenes de ecografías de tiroides todas previamente recortadas omitiendo información sensible de cada paciente para efectos de normatividad.



Figura 3- Imagen en formato original

Con los datos listos para entrenar el sistema, se construye una red neuronal que tiene como característica principal la extracción de características, para ello fue necesario como método principal la investigación de los diferentes tipos de redes neuronales y sus usos principales. Esta investigación da como resultado que las redes neuronales convolucionales se enfocan en el procesamiento de imágenes para su clasificación donde se aplica la prueba de tamizaje que se divide en 5 tipos, que aplicados a nuestra red neuronal convolucional son las salidas u outputs que generaron una sensibilidad por validación realizada y las pérdidas del sistema.

3.1. CONSTRUCCIÓN DE LA RED NEURONAL COVOLUCIONAL

Se distribuyen las imágenes de desde TIRADS 1 a TIRADS 5 en tres carpetas (train, test, val) entrenamiento, prueba y validación. La carpeta de entrenamiento contiene mil imágenes, la de prueba 623 imágenes y la de validación un promedio de 25 imágenes por carpeta de TIRADS 1 a TIRADS 5. Con lo anterior listo, se realizó el algoritmo de predicción con 4 convoluciones, 4 Max Pooling, 1 flatland, dos capas conectadas con 512 neuronas y 2 dropout. Para el entrenamiento del modelo se aplicaron 10 épocas y con 163 pasos y 16 bloques.

3. METODOLOGÍA PARA EL DESARROLLO

Para proporcionar una descripción normalizada del ciclo de vida de nuestro proyecto para el análisis de datos, se da uso del modelo CRISP-DM [4] que cubrió todas las fases del desarrollo, donde abarcamos las siguientes pautas para la investigación, desarrollo y despliegue:

I. Comprensión del negocio o del contexto del proyecto (objetivos y requerimientos desde una perspectiva no técnica) [5]:

Para el proceso de identificación del área a investigar dando como resultado la solución a una problemática en el área médica, fue fundamental la orientación y guía dada por especialistas de la clínica FOSCAL y miembros del semillero de radiología de la UNAB, quienes dieron a conocer por medio de evidencia y estudios que se han realizado en el país, que dan como resultado una alta demanda en lecturas de imágenes para identificar posibles anomalías en la glándula tiroidea por medio de ecografías, donde los especialistas en radiología son los actores principales, quienes generan un diagnóstico para cada paciente.

Esta área de la medicina presenta una creciente necesidad en la implementación de un sistema de inteligencia artificial que promedio de un algoritmo de predicción haga una clasificación en ecografías de tiroides haciendo uso de la prueba de tamizaje EU-TIRADS, para dar un apoyo fundamental a la hora de tomar decisiones en el diagnóstico, con el fin de mejorar el servicio de salud en el área de radiología.

II. Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio) [6]:

Se tuvo en cuenta como criterio de selección de datos, ecografías de la glándula tiroidea en un formato plano, sin marcaciones, sin medidas y sin el efecto Doppler como procesamiento inicial, que cumplan con los objetivos específicos de la minería de datos donde enfocamos el aprendizaje en la glándula que se estudia, puesto que el dataset que se diseñó en su mayoría eran imágenes de este tipo. Para un mejor aprendizaje de la máquina es necesario que los datos estén unificados y que no varíen para no forzar y sesgar el aprendizaje.

Se elaboró una tabla en Excel para facilitar el proceso de clasificación que realizó el doctor Federico Lubinus tomando como herramienta la prueba de tamizaje EU-TIRADS, la misma con la cual se entrenó el sistema. Esto se hace con el fin de organizar las imágenes que le serán generadas en el proceso de entrenamiento, pruebas y validación.

Todas estas imágenes que se clasificaron fueron generadas en su formato original DICOM (Imagen digital y comunicación sobre medicina), formato de imagen médica, con el que fue necesario implementar un script para convertirlo a png, con el fin de hacer un procesamiento más ligero.

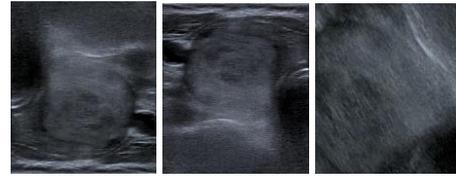


Figura 4 Selección de imágenes. Fase inicial (imagen 1 flip, imagen 2 rotation, imagen 3 random Crop)

Para el entrenamiento fue necesario diseñar la base de datos con las imágenes generadas por la clínica FOSCAL, después de este proceso se eliminan los datos sensibles para evitar procesos legales. Se recortaron las imágenes por medio de un script.

III. Preparación de los datos para el procesamiento (Selección de técnica de modelado) [7]:

Se realiza una serie de investigaciones acerca de los algoritmos de predicción con mayor eficiencia en imágenes, para ello se evalúan los diferentes tipos de redes neuronales que se encargan del procesamiento y sus funcionalidades. Se toma como algoritmo de predicción en imágenes las redes neuronales convolucionales o CNN.

Se construye el modelo de red neuronal de acuerdo con el data que le generamos. Para una mejor sensibilidad en los resultados en necesario una base de datos de gran cantidad de imágenes clasificadas previamente para su entrenamiento con ello el resultado de sensibilidad será óptimo.

Se normalizan todas las imágenes de tal manera que queden de un mismo formato, ya que son generadas en DICOM. Al tener las imágenes archivadas en carpetas y clasificadas, se hace el entrenamiento de nuestra red neuronal, por medio de 4 convoluciones, 4 Max pooling, 1 flatten, dos capas conectadas con 512 neuronas y 2 dropout. Para el entrenamiento del modelo se aplicaron 10 épocas y cada uno con 163 paso de 16 bloques.

IV. Modelado (Aplicamos las técnicas de minería de datos a los dataset) [8]:

Basados en una investigación reciente que se realizó, según pruebas que se hicieron para predicciones en imágenes, se propone un nuevo algoritmo de entrenamiento que corrige las desventajas de CNN, así mismo mejora su velocidad y precisión. Llamamos a este método Fast R-CNN que es comparativamente rápido para entrenar y validar un sistema con procesamiento en imágenes.

Con este algoritmo se espera un accuracy por encima de 70% debido a que es moderadamente bueno para hacer validaciones y mejorarlo si es el caso de ser implementado.

V. Evaluación (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio) [9]:

Se implementaron 3 modelos de evaluación para el algoritmo de clasificación, estos fueron: accuracy, matriz de confusión y F1, estos evalúan la precisión de los resultados y las pérdidas que tuvo en las pruebas después de realizar el entrenamiento. Estas métricas consisten en:

- **Accuracy:** Es la relación entre el número de predicciones correctas y el número total de las muestras de entrada. Da una buena precisión si hay un número igual de muestras que pertenecen a cada clase.

$$Accuracy = \frac{\text{Number of Correct predictions}}{\text{Total number of predictions made}}$$

- **Matriz de confusión:** Da como resultado una matriz y describe el rendimiento completo del modelo. Para ello se tienen en cuenta cuatro términos importantes:

1. Verdaderos positivos
2. Negativos verdaderos
3. Falsos positivos
4. Falsos negativos

$$ConfusionMatrix = \frac{\text{TruePositives} + \text{FalseNegatives}}{\text{TotalNumberofSamples}}$$

- **Puntaje de F1:** Se usa para medir la presión de una prueba, F1 es la media armónica entre precisión y recuperación, su rango es de [0,1]. Su resultado nos indica cuántas instancias clasifica correctamente y qué tan robusto es debido a que no pierde número significativo de instancias.

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Para que sea un resultado óptimo y con fines de implementación futura es necesario que los resultados de la evaluación estén por encima de un 70%, verificar el porcentaje de pérdida y mejorar el entrenamiento para aumentar la precisión, son propuestas que se deben tener en consideración. Un buen resultado depende de la cantidad de imágenes del dataset de entrenamiento, por ello se tiene una base de datos con más de 1.000 imágenes.

Realizar un despliegue del sistema depende de los procedimientos legales que se manejan en cada hospital y en el país debido a que dar un diagnóstico

asistido por una máquina no está considerado en las normas, debido a que es obligatorio que cada hospital o clínica tenga en su planta un médico especialista en radiología quien se encarga de la lectura de imágenes.

Con ello se espera realizar pruebas para comparar resultados entre el sistema de AI y el diagnóstico de un radiólogo. Este proceso se realiza con dos especialistas de diferentes categorías, uno senior y otro junior, se comparan resultados para evaluar la especificidad y sensibilidad de la máquina.

4. RESULTADOS DE LA PRUEBA INICIAL

Se realizó una prueba inicial con 1.510 imágenes por cada TIRADS, en este caso son 4 carpetas, para el entrenamiento. El hospital generó 256 imágenes clasificadas según el EU-TIRADS, estas fueron evaluadas por el doctor Federico Lubinus quien nos hizo entrega de un formato en Excel, donde cada imagen tiene un código de identificación y la clasificación de TIRADS 2 a 5.

Como requerimiento para una buena predicción se requirió aumentar de la Base de Datos con la codificación de un script para el DataAugmentation, donde se usaron (4) tipos de técnicas: Volteo (Flip), rotación (Rot), Corte aleatorio (Random Crop) y traslación (Translation). Los resultados iniciales generados por el sistema fueron los siguientes:

```

=====
=
Total params: 1,510,436
Trainable params: 1,510,436
Non-trainable params: 0
=====
-
The loss is: 0.7854722690582275 .
The testing accuracy is : 65.0 %
#####PREDICTION#####
200
#####CONFUSION MATRIX#####
[[ 2 25 12 11]
 [ 4 31  8  7]
 [ 5 23 11 11]
 [ 6 22 15  7]]
#####CLASSIFICATION REPORT#####
          precision    recall  f1-score   support

 TIRADS2     0.12     0.04     0.06         50
 TIRADS3     0.31     0.62     0.41         50
 TIRADS4     0.24     0.22     0.23         50
 TIRADS5     0.19     0.14     0.16         50

 accuracy          0.26         200
 macro avg         0.21         200
 weighted avg      0.21         200

(tesis_estefania) bash-3.2$

```

Se asignaron 4 de capas convolucionales, 4 de capas de “pooling”, 512 neuronas de capas conectadas una capa de “flatten” y finaliza con (5) perceptrones multicapa para la clasificación. Se hace un procesamiento con 10 épocas, 163 pasos y 16 bloques, que dan como resultado un accuracy de 65%, una pérdida de 0,78, se hizo una predicción de 200 imágenes.

5. CONCLUSIONES

Como conclusión se comprueba la utilidad del uso de redes neuronales convolucionales para la clasificación de imágenes de cualquier tipo, incluyendo las imágenes médicas que por su alto nivel de detalle y complejidad pueden representar un reto para otros tipos de modelos de aprendizaje computacional. Es importante realizar pruebas con diferentes tipos de redes neuronales, actualmente existe un algoritmo llamado “Fast R-CNN” y “R-CNN” que tiene un mejor rendimiento en el procesamiento y aprendizaje, por ello se ponen en consideración para realizar futuros proyectos o mejorar el rendimiento de este.

El primer gran reto para sortear cuando se trabaja la clasificación de imágenes médicas está en conseguir un dataset apropiado, con muestras suficientes para cada una de las categorías, validado por expertos y del cual se puedan extraer las características principales necesarias para el aprendizaje. Finalmente, después de realizar las métricas del modelo, se concluye que los valores de la pérdida logarítmica y el accuracy por sí solos no son suficientes para evaluar la efectividad del modelo por esta razón se implementan otras métricas de evaluación que dan como resultado un bajo porcentaje en el cual se requiere seguir trabajando para mejorar el modelo debido al sesgo de los datos.

REFERENCIAS

- [1] Pardo C, Cendales R. Incidencia, mortalidad y prevalencia de cáncer en Colombia 2007-2011. 1, 1-147. 2015. Bogotá D.C., Instituto Nacional de Cancerología
- [2] Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, Maranghi M, Falcone R, Ramundo V, Cantisani V, Filetti S, Durante C. Reducing the Number of Unnecessary Thyroid Biopsies While Improving Diagnostic Accuracy: Toward the "Right" TIRADS. *J Clin Endocrinol Metab.* 2019 Jan 1;104(1):95-102. doi: 10.1210/jc.2018-01674.
- [3] © 2017 Instituto Nacional de Cancerología ESE. (s.f.). [Análisis de Situación del Cáncer en Colombia 2015]. Recuperado de https://www.cancer.gov.co/Situacion_del_Cancer_en_Colombia_2015.pdf
- [4] Singular - CRISP-DM: La metodología para poner orden en los proyectos de Data Science. Available, 8 de agosto de 2016 Recuperado de: <https://data.sngular.com/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science>.
- [5] Data.sngular.com. (2019). Singular - CRISP-DM Fase I. Comprensión del negocio (Business Understanding). Available at: <https://data.sngular.com/es/art/26/crisp-dm-fase-i-comprension-del-negocio-business-understanding>.
- [6] Data.sngular.com. (2019). singular - CRISP-DM Fase II. Data Understanding. Estudio y comprensión de los datos. Recuperado de: <https://data.sngular.com/es/art/27/crisp-dm-fase-ii-data-understanding-estudio-y-comprension-de-los-datos>.
- [7] Data.sngular.com. (2019). singular - CRISP-DM Fase III. Data Preparation. Análisis de los datos y selección de características. Recuperado de: <https://data.sngular.com/es/art/28/crisp-dm-fase-iii-data-preparation-analisis-de-los-datos-y-seleccion-de-caracteristicas>.
- [8] Data.sngular.com. (2019). *sngular - CRISP-DM Fase IV. Modeling. Modelado.* Recuperado de: <https://data.sngular.com/es/art/29/crisp-dm-fase-iv-modeling-modelado>.
- [9] Data.sngular.com. (2019). singular - CRISP-DM Fase V. Evaluation. Evaluación (obtención de resultados). Recuperado de: <https://data.sngular.com/es/art/30/crisp-dm-fase-v-evaluation-evaluacion-obtencion-de-resultados>.
- [10] Cv Fundación (2015). Fast R-CNN. IEEE. Recuperado: https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf.