Universidad de los Llanos

ORINOQUIA
Suplemento Volumen 25 - Número 1, Enero-Junio 2021

Artículo Original/Original Article

Un análisis comparativo de arquitecturas de *Redes Neuronales Convolucionales* para clasificación de grado de riesgo de lesiones de próstata a partir de imágenes unimodales o bimodales de *mpMRI*

A comparative analysis of Convolutional Neural Network architectures for classifying the degree of risk of prostate lesions from unimodal or bimodal *mpMRI* images

Uma análise comparativa das arquiteturas de rede neural convolucional para classificar o grau de risco de lesões da próstata a partir de imagens unimodais ou bimodais de *mpMRII*

Recibido: 24 de noviembre de 2020. **Aceptado**: 17 de mayo de 2021.

***Mauricio Caviedes-Rojas*[1],**
Ing. Sist; https://orcid.org/0000-0002-1149-8508

***Charlems Alvarez-Jiménez*[2],**
Ing. Sist, MSc; https://orcid.org/0000-0001-7100-6387

***Eduardo Romero-Castro*[3],**
Méd. Ciruj, MSc, PhD;
https://orcid.org/0000-0003-2088-2509

***Angel Cruz-Roa*[4],**
Ing. Sist, MSc, PhD; https://orcid.org/0000-0003-3389-8913

1 Grupo de investigación GITECX & AdaLab, Universidad de los Llanos, Villavicencio, Colombia.
Email: Jerson.caviedes@unillanos.edu.co
2 Grupo de investigación CIM@LAB, Universidad Nacional de Colombia, Bogotá D.C., Colombia.
Email: calvarezj@unal.edu.co
3 Grupo de investigación CIM@LAB, Universidad Nacional de Colombia, Bogotá D.C., Colombia.
Email: edromero@unal.edu.co
4 Grupo de Investigación GITECX & AdaLab, Universidad de los Llanos, Villavicencio, Colombia.
Email: aacruz@unillanos.edu.co

Suplemento Orinoquia, Enero-Junio 2021; 25(1):39-55
ISSN electrónico: 2011-2629
ISSN impreso: 0121-3709
https://doi.org/ 10.22579/20112629.683

Resumen

Este trabajo presenta un análisis comparativo de cinco arquitecturas de redes neuronales convolucionales (*CNN*) usando imágenes de resonancia magnética multiparamétrica (*mpMRI*) para la clasificación de tejidos con presencia de lesiones de cáncer próstata. Como conjunto de datos de entrenamiento y validación se usó *SPIE-AAPM-NCI Prostate MR Classification Challenge*, el cual cuenta con 344 casos de imágenes de resonancia magnética de las modalidades: *T2W* (T2 Ponderado), *ADC* (Coeficiente de Difusión Aparente) y *Ktrans* (imágenes preprocesadas de la modalidad *DCE* - Dinámico de Contraste Mejorado), a partir del cual se usaron tres subconjuntos de datos de una sola modalidad independiente (unimodal): *T2W*, *ADC* y *Ktrans*, y dos subconjuntos de datos combinando dos modalidades (bimodal): *Ktrans-ADC* y *Ktrans-T2W*, para su comparación y análisis. A partir de la escala de Gleason (*Gleason score - GS*) y el grado ISUP (*International Society of Urologic Pathologists*), las cuales son usada para medir el grado de agresividad del cáncer de próstata, se establecieron dos niveles de agresividad: Bajo y Alto. La clase Bajo son aquellas lesiones con GS = 6, y la clase Alto, las lesiones con el valor del GS $\geq$ 7. Los resultados experimentales muestran un rendimiento superior con las imágenes de la modalidad *Ktrans* en las 4 primeras arquitecturas obteniendo un valor máximo de AUC (*area under ROC curve* o área bajo la curva) de 0.71 $\pm$ 0.127. Sin embargo, la quinta arquitectura inspirada en la *LetNet* combinando dos modalidades de *mpMRI*, Ktrans-T2W, se obtiene un AUC de 0.72 $\pm$ 0.058, lo cual sugiere ligeramente que, aunque la modalidad *Ktrans* es la más relevante, su combinación con *T2W* podría mejorar la precisión diagnóstica.

Palabras clave: cáncer de próstata, resonancia magnética multiparamétrica, redes neuronales convolucionales.

Abstract

This work presents a comparative analysis of five convolutional neural network (CNN) architectures using multiparametric magnetic resonance imaging (mpMRI) for the classification of tissues with the presence of prostate cancer lesions. SPIE-AAPM-NCI Prostate MR Classification Challenge was used as a training and validation data set, which consists of 344 cases with magnetic

Como Citar (Norma Vancouver):
Caviedes-Rojas M, Alvarez-Jiménez CA, Romero-Castro E, Cruz-Roa A. Un análisis comparativo de arquitecturas de Redes Neuronales Convolucionales para clasificación de grado de riesgo de lesiones de próstata a partir de imágenes unimodales o bimodales de mpMRI. Orinoquia, 2021;(SUPLEMENTO 1):39-55. Https://doi.org/ 10.22579/20112629.683

resonance images from the modalities: T2W (Weighted T2), ADC (Apparent Diffusion Coefficient), and Ktrans (preprocessed images from the DCE -Dynamic Enhanced Contrast- modality), from which three subsets of data from a single independent modality were used (unimodal): T2W, ADC and Ktrans, and two subsets of data combining two modalities (bimodal): Ktrans-ADC and Ktrans-T2W, for comparison and analysis. From the Gleason scale (Gleason score - GS) and the ISUP grade (International Society of Urologic Pathologists), which are used to measure the degree of aggressiveness of prostate cancer, two levels of aggressiveness were established: Low and High. The Low class is those lesions with GS = 6, and the High class, those lesions with the GS value ≥ 7. The experimental results show a superior performance with Ktrans modality images in the first 4 architectures obtaining a maximum AUC value (area under ROC curve) of 0.71 ± 0.127. However, the fifth LetNet inspired architecture combining two mpMRI modalities, Ktrans-T2W, obtains an AUC of 0.72 ± 0.058, which slightly suggests that although the Ktrans modality is the most relevant, its combination with T2W could improve diagnostic accuracy.

Keywords: Prostate Cancer, Multiparametric Magnetic Resonance Imaging, Convolutional Neural Networks.

Resumo

Este trabalho apresenta uma análise comparativa de cinco arquiteturas de redes neurais convolutivas (CNN) usando ressonância magnética multiparamétrica (mpMRI) para a classificação dos tecidos com a presença de lesões do câncer de próstata. SPIE-AAPM-NCI Prostate MR Classification Challenge foi usado como um conjunto de dados de treinamento e validação, que conta com 344 casos com imagens de ressonância magnética das modalidades: T2W (T2 ponderado), ADC (Coeficiente de difusão aparente) e Ktrans (imagens pré-processadas da modalidade DCE (Dynamic Enhanced Contrast)), dos quais foram usados três subconjuntos de dados de uma única modalidade independente: T2W, ADC e Ktrans, e dois subconjuntos de dados combinando duas modalidades: Ktrans-ADC e Ktrans-T2W, para comparação e análise. Da escala Gleason (Gleason score - GS) e da ISUP (International Society of Urologic Pathologists), que são usadas para medir o grau de agressividade do câncer de próstata, dois níveis de agressividade foram estabelecidos: Baixo e Alto. A classe Baixa são aquelas lesões com GS = 6, e a classe Alta, aquelas lesões com valor GS ≥ 7. Os resultados experimentais mostram um desempenho superior com as imagens da modalidade Ktrans nas primeiras 4 arquiteturas obtendo um valor máximo de AUC (área sob curva) de 0,71 ±0,127. Entretanto, a quinta arquitetura inspirada na LetNet combinando duas modalidades mpMRI, Ktrans-T2W, obtém uma AUC de 0,72 ±0,058, o que sugere ligeiramente que, embora a modalidade Ktrans seja a mais relevante, sua combinação com T2W poderia melhorar a precisão do diagnóstico.

Palavras-chave: Câncer de próstata, ressonância magnética multiparamétrica, redes neurais convolucionais.

Introducción

La resonancia magnética (RM) es una tecnología ampliamente usada en medicina para la detección de enfermedades y el monitoreo de tratamientos. La RM consiste en un proceso de tomografía de emisión que se basa en la excitación de los núcleos de los átomos de hidrógeno. La RM tiene varias ventajas: capacidad de adquisición multiplanar, una elevada resolución de contraste, ausencia de efectos nocivos por radiación ionizante y una elevada resolución de contraste (Lafuente Martínez, Luis, y Moreno, 2016).

Según GLOBOCAN (*Global Cancer Observatory*), para el 2018 alrededor del mundo se presentaron 1'276,106 de nuevos casos de cáncer de próstata, donde para ese mismo año se produjeron 358,989 muertes relacionados con este tipo de cáncer. En cuanto a Colombia en el 2018, GLOBOCAN registró 12,712 de nuevos casos de cáncer de próstata y el número de personas que fallecieron fue de 3,166 (*GLOBOCAN*, 2018). Algunos factores de riesgos que pueden causar cáncer de próstata son la edad, donde existe una alta probabilidad que un hombre después de los 50 años pueda padecerlo, y de hecho, alrededor de 6 de cada 10 casos de cáncer de próstata se detectan en hombres mayores de 65 años. También influye la raza o el grupo étnico, por ejemplo, en los hombres de raza negra y con ascendencia africana, el cáncer de próstata ocurre con más frecuencia, además del factor genético debido a los antecedentes familiares (*American Cancer Society*, 2019).

En las últimas décadas se ha trabajado para mejorar la tecnología diagnóstica, en especial en estadios tempranos. Es por eso que hace unos años se ha abordando la resonancia magnética multiparamétrica (mpMRI), en particular para diagnosticar cáncer de próstata. La resonancia magnética multiparamétrica de próstata consiste en la combinación de imágenes anatómicas de alta resolución tradicionales (por ejemplo, T2) con técnicas de imágenes funcionales, por ejemplo, resonancia por difusión ponderadas (*DWI* por sus siglas en inglés), resonancia magnética por perfusión (*PRM*) o espectroscopía. Las imágenes *T2W* (secuencia T2 ponderadas) se usan para analizar la morfología de la zona prostática con el fin de evaluar anomalías en la zona de transición (TZ), mientras las imágenes *DWI* miden los movimientos de las moléculas de agua y se usan para la detección de cáncer en la zona prostática periférica (Barentsz *et al*., 2016). En la actualidad estos

avances han ido aún más allá, al introducir la inteligencia artificial en el campo de la medicina para contribuir en la mejora de los diagnósticos en el cáncer de próstata. Por ejemplo, existen trabajos donde han aplicado diferentes métodos de clasificación, como el estudio de Fehr *et al*., (Fehr, 2015), donde se utilizó una *SVM* (Máquina de soporte vectorial) para la tarea de clasificar entre cáncer y no cáncer de próstata usando solo la modalidad *ADC* alcanzando una exactitud media de 0.82 para la zonas periférica y transicional (PZ y TZ) y una exactitud media de 0.84 para la zona periférica (PZ) solamente, sin realizar estrategia de aumento de datos a partir de características basadas en textura de la matriz de co-ocurrencias de niveles de gris (GLCM) sobre regiones cuidadosamente segmentadas de las lesiones. Al incluir la modalidad T2 a ADC no mejoraba el desempeño y al incluir métodos de sobremuestreo incluso empeoraba. Los mejores resultados de Fehr *et al*., lo alcanzaron usando una *Recursive Feature Selection Support Vector Machine* (RFE-SVM) con exactitud media entre 0.83 y 0.93 en conjunto con estrategias de sobremuestreo (SMOTE o Gibbs), sin embargo, puede estar tendiendo al sobreajuste, sin olvidar que requiere un trabajo manual de una segmentación precisa de la lesión. La investigación de Wibmer *et al*. (Wibmer *et al*., 2015) realizó un análisis de las características de textura de Haralick para valorar su utilidad en la diferenciación de tejidos no cancerosos o cancerosos de diferentes grados de Gleason para cáncer de próstata. Las características de textura de Haralick extraídas de las imágenes fueron: energía, entropía, correlación, homogeneidad e inercia, y un análisis estadístico usando estadística descriptiva y las diferencias entre los valores de estas para lesiones cancerosas y no cancerosas calculado con ecuaciones de estimación generalizadas (GEE), una matriz de covarianza robusta y una estructura de correlación independiente, permitió determinar una mejor diferenciación estadísticamente significativa para la zona periférica (PZ) de forma independiente para cada una de las modalidades *T2W* y *ADC*, y mejor capacidad de tres características de Haralick de la modalidad ADC para diferenciación del grado de Gleason, siempre que la región esté previamente segmentada en ambas tareas. Tiwari *et al*., (Tiwari, Kurhanewicz, y Madabhushi, 2013) implementaron un método de aprendizaje semi-supervisado basado en *Multi-kernel graph embedding* (*SeSMiK-GE*) sobre un enfoque *leave-one-out* en 29 estudios que alcanza un valor de promedio de *AUC* de 0.89, mientras que para T2w MRI fue de 0.54, para espectroscopia de resonancia magnética (MRS) fue de 0.61, y concatenando las características de las dos modalidades (T2w MRI y MRS) alcanzó 0.64. Por último, en el trabajo de Le *et al*., (Le, 2017) se presenta una red neuronal convolucional (*CNN*) multimodal, la cual fusiona las dos modalidades (ADC y T2W) integrando características manuales extraídas de cada una usando una nueva función de pérdida de similitud, donde los resultados de la *CNN* multimodal son combinados con los resultados basados en características manuales del estado del arte utilizando una *SVM*, alcanzando una sensibilidad 89.85% y una especificidad 95.83% para distinguir entre cáncer y no cáncer de próstata de datos de 364 pacientes. Sin embargo, cada trabajo previamente mencionado usa un conjunto de datos de fuentes diferente, con cantidades distintas de casos y modalidades de imágenes de mpMRI en cada uno, con variadas estrategias de combinación y diseños experimentales de evaluación diferentes.

Las redes neuronales convolucionales (CNN) se han utilizado durante décadas en el campo de visión por computador, aunque sólo fue conocido su verdadero valor en la competición *ImageNet* en el 2012, donde el uso eficiente de las unidades de procesamiento gráfico (GPU), el aumento de la cantidad de datos, además del uso de nuevas técnicas permitieron que se convirtiera en uno de los mayores avances en el campo del aprendizaje automático en visión por computador (Tajbakhsh *et al*., 2016). Tradicionalmente, los modelos de aprendizaje automático eran entrenados para realizar tareas basadas en la extracción manual de características de los datos sin procesamiento. En cambio, en el aprendizaje profundo (*Deep Learning* en inglés), los algoritmos obtienen modelos que aprenden representaciones y características eficientes de manera automática, tomadas directamente de los datos (Lundervold y Lundervold, 2019). Existen diferentes métodos de aprendizaje profundo para la aplicación en diferentes ámbitos o tareas, por ejemplo, en el procesamiento del lenguaje natural, procesamiento de imágenes hiperespectrales y análisis de imágenes médicas (Tajbakhsh *et al*., 2016). Las *CNN* como método de aprendizaje profundo ha alcanzado una posición importante en el campo de visión por computador y ha destacado por su alto desempeño para la clasificación de imágenes. Por lo anterior, considerando algunos trabajos de *Deep Learning* usando *CNN* en el contexto de imágenes médicas incluyendo modalidades de resonancia magnética, es de interés un análisis comparativo con un mismo conjunto de datos y diseño experimental para establecer el aporte independiente usando únicamente modelos de *Deep Learning* basados en *CNN* para la clasificación de lesiones de próstata por grado

de riesgo (baja o alta agresividad) más que solamente entre tejido canceroso y no canceroso usando diferentes modalidades de mpMRI o combinaciones de estas, incluyendo la modalidad *Ktrans*.

Existen algunos pocos trabajos previos basados en CNN sobre un mismo conjunto de datos (PROSTATEx Challenge) para su comparación y análisis apropiado. El trabajo de Saifeng *et al*., (Saifeng, 2017) obtuvo un *AUC* de 0.84, mientras que el de Alireza *et al*., (Alireza, 2017), alcanzó un *AUC* de 0.80, y más recientemente el trabajo de Gutiérrez *et al*., (Gutiérrez, 2020) obtuvo un *AUC* promedio de 0.72 al combinar 4 modalidades de mpMRI (*T2 transaxial, T2 sagital, ADC, Ktrans*) y 0.74 al explorar la ponderación de pesos (entre 0 a 1) entre dos modalidades (*T2 transaxial, Ktrans*), siendo de 0.8 el peso mayor para la modalidad *Ktrans*. Estos trabajos mencionados anteriormente, se enfocaron en la tarea de clasificar las lesiones de cáncer de próstata. Pero también se implementan las CNN para tareas de segmentación utilizando el mismo conjunto de datos para especificar la región con presencia de cáncer de próstata, como es el caso del trabajo de Yongkai *et al*., (Yongkai, 2017) obteniendo un AUC entre 0.74 – 0.86. En la Tabla 1 puede ver la información completa sobre la comparación con estos trabajos.

El objetivo de este trabajo es hacer un análisis comparativo de algunas arquitecturas diferentes de *Deep Learning* basadas en Redes Neuronales Convolucionales para la clasificación de lesiones de la próstata de las zonas PZ y TZ a partir de imágenes digitales unimodales o bimodales de resonancia magnética multiparamétrica (T2W, ADC y *Ktrans*). Los resultados que se obtuvieron alcanzaron un desempeño de área bajo la curva ROC (AUC) promedio de 0.71 ± 0.127 con el uso de la modalidad *Ktrans* (unimodal), y un valor de AUC promedio de 0.72 ± 0.058 combinando las modalidades *Ktrans-T2W* (bimodal) en un diseño experimental de validación cruzada de *5-fold*.

Materiales y métodos

Conjunto de datos de mpMRI de cáncer de próstata

El conjunto de datos que se utilizó en el desarrollo de ese trabajo proviene del desafío *PROSTATEx Challenge* ("*SPIE-AAPM-NCI Prostate MR Classification Challenge*"). Las imágenes que se proporcionan corresponden a las modalidades: *T2W, Ktrans* (imágenes post-procesadas de la modalidad *DCE*) y *ADC*. El objetivo principal de este reto consiste en la clasificación de la agresividad de lesiones de cáncer de próstata para esto se definen dos tipos de clases, clínicamente significativo (esto significa la presencia o desarrollo de cáncer agresivo) e indolente (cáncer con agresividad baja). En la Tabla 2 se describen las clases utilizadas en este proyecto en términos de la escala Gleason y grado ISUP como los dos tipos de lesiones: alta y baja agresividad. La mayoría de los casos tiene en cada modalidad su respectiva imagen con la ubicación espacial de la lesión. Este conjunto de datos se puede encontrar disponible en el siguiente enlace https://prostatex.grand-challenge.org/.

El conjunto de datos cuenta con 344 casos de estudio, y para cada caso existe la lesión representada en las tres modalidades (*T2W, ADC y Ktrans*). La distribución del conjunto de datos por número de casos se encuentra detallada en la Tabla 3. El subconjunto de entrenamiento se usó para construir el modelo, dado que las lesiones se encontraban anotadas previamente. La Tabla 4 describe la distribución de los grupos de las lesiones según la clase por nivel de riesgo (bajo o alto). En la Figura 1 se puede visualizar un ejemplo de una lesión de cáncer de próstata en las tres modalidades. La resolución de las imágenes varía según la modalidad, como se describe en la Tabla 5.

Finalmente, para el análisis comparativo se crearon dos nuevos conjuntos de datos de imágenes "bimodales" (unión de dos modalidades), un conjunto datos lo

Tabla 1. Comparación de desempeños con trabajos previos que utilizaron el mismo conjunto de datos.

Autor	Método de clasificación	Descripción	AUC
(Saifeng , 2017)	CNN	3D Multiparametric MRI	0.84
(Alireza , 2017)	CNN	3D Multiparametric MRI	0.80
(Jarrel , 2017)	CNN	-	0.84
(Yongkai., 2019)	CNN	Segmentation	0.74 – 0.86
(Gutiérrez., 2020)	CNN	-	0.74

Tabla 2. Descripción de las clases en la clasificación según la escala de Gleason y el grado ISUP.

Clase	Escala Gleason	Grado ISUP
Baja o Indolente	2-6	Grado 1
Alto o Clínicamente Significativo	3+4 = 7	Grado 2
	4+3 = 7	Grado 3
	4+4 = 8 3 +5 = 8 5+3 = 8	Grado 4
	9-10	Grado 5

Tabla 3. Distribución del conjunto de datos por casos en cada modalidad.

	T2W	ADC	Ktrans
Entrenamiento	204	204	204
Prueba	140	140	140

Tabla 4. Distribución por clase en subconjunto de entrenamiento.

Entrenamiento	Alto	Bajo
T2W	76	244
ADC	76	244
Ktrans	76	244

Tabla 5. Descripción de las dimensiones de las imágenes en las modalidades.

Modalidad	Dimensiones
T2W	320x320x19 320x320x21 384x384x19 384x384x21 384x384x23 384x384x25 640x640x21
ADC	84x128x19 84x128x20 84x128x21 84x128x23 84x128x25 120x128x19 120x128x20 128x106x20
Ktrans	128x128x12 128x128x16

componen las modalidades *Ktrans-T2W* y el segundo *Ktrans-ADC* (ver Figura 2).

Preprocesamiento del conjunto de datos

Para el preprocesamiento del conjunto de datos se implementaron dos métodos, la normalización de Min-Max y normalización estándar. En la normalización de Min-Max el objetivo es que los valores de los píxeles se establezcan en un intervalo entre 0 y 1, esto porque los valores de intensidad de los píxeles en las imágenes son variables y en el caso de este conjunto los valores

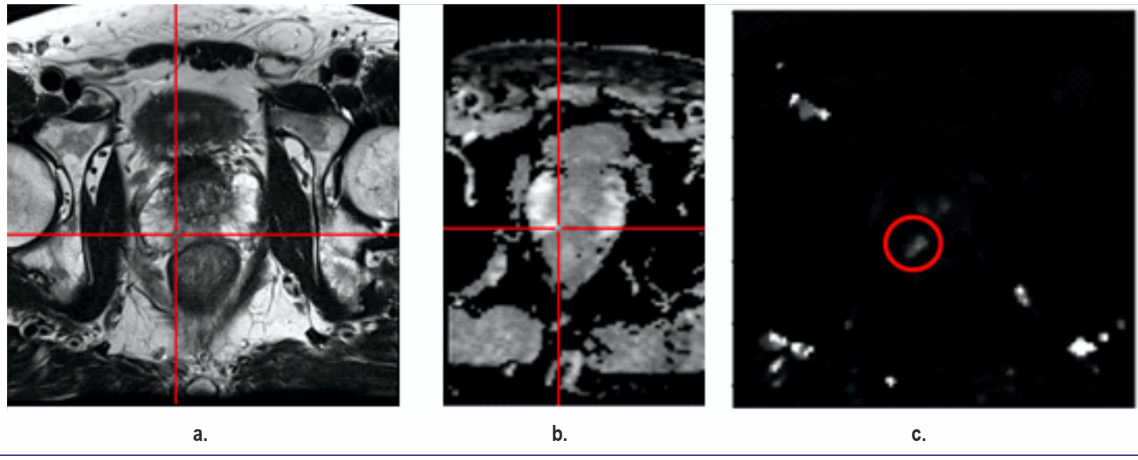

Figura 1. Ejemplos de las modalidades que componen el conjunto de datos, a. T2W, b. ADC y c. Ktrans (imágenes post-procesadas de la modalidad DCE).

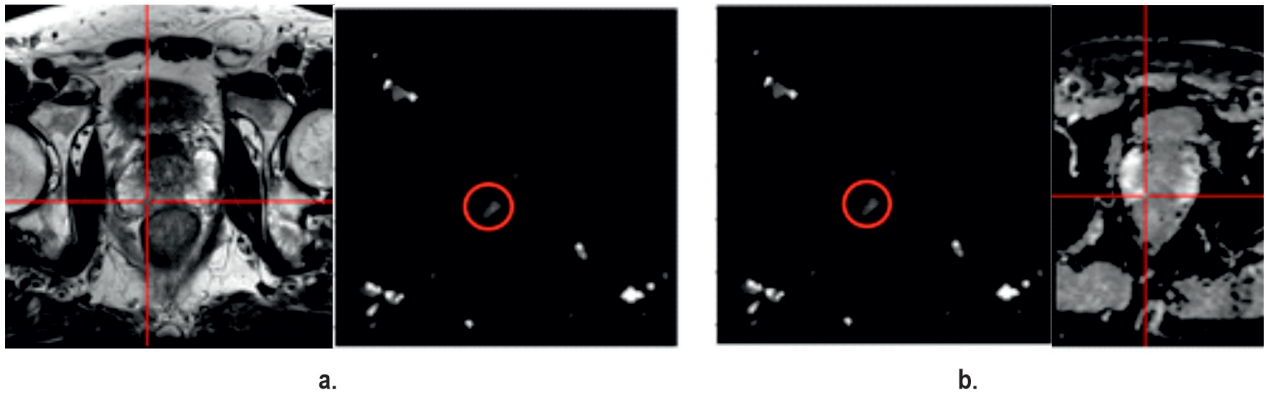

Figura 2. Ejemplo del conjunto de datos bimodales (a) Ktrans- T2W y (b) Ktrans-ADC

de los píxeles están en el rango de 0 a 65,536. La normalización Min-Max se describe en la Ecuación 1.

$$X' = \frac{(X - X_{min})(1 - 0)}{(X_{max} - X_{min})} \quad (1)$$

- ***X***, es el valor del píxel original que se va a ajustar en el rango de 0 a 1.
- ***Xmin***, es el valor mínimo del píxel que pertenece a la imagen.
- ***Xmax***, es el valor máximo del píxel que pertenece a la imagen.
- ***X'***, es el valor del píxel ajustado en el rango de 0 a 1.

Por otro lado, la normalización estándar consiste en modificar los valores de los píxeles de las imágenes para que tengan media cero y varianza uno, como se describe en la Ecuación 2.

$$X' = \frac{X - \mu}{\sigma} \quad (2)$$

- ***X***, valor del píxel original que se va a normalizar.
- μ, el valor de la media de los píxeles que componen la región de la imagen.
- σ, el valor de la desviación estándar de los píxeles que componen la región de la imagen.
- ***X'***, es el valor del píxel normalizado.

En este trabajo, se extrajo para cada caso y en cada modalidad (*ADC*, *T2W* y *Ktrans*), una región de interés de tamaño 32x32x3 píxeles, donde esta región o "parche" tiene en su centro la lesión de interés de diagnóstico. Para conformar los conjuntos de las imágenes bimodales (*Ktrans-ADC* y *Ktrans-T2W*) se tomó la imagen que contiene la lesión de los parches generados en las modalidades (*ADC, T2W y Ktrans*) y luego se unen, generando parches de 32x32x2 píxeles, conformando así matrices de 3 dimensiones. Dado que las imágenes de *ADC* y *T2W* tienen diferentes resoluciones, como se

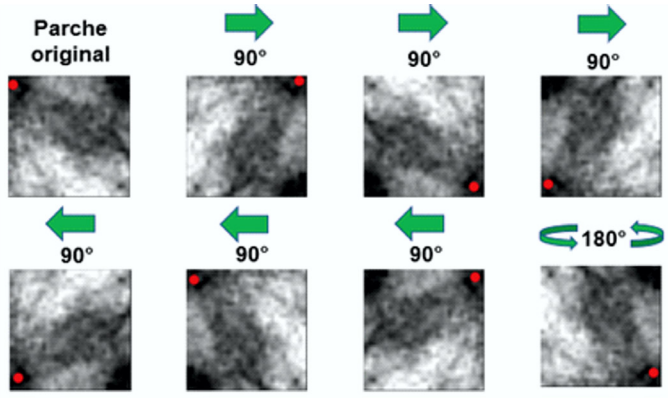

Figura 3. Implementación de la técnica sobremuestreo para aumentar el conjunto de datos de entrenamiento.

observa en la Tabla 5, fue necesario hacer un redimensionamiento (reescalado) de algunas de las imágenes y así manejar una resolución uniforme para todas. En el caso de la modalidad de ADC, se redimensionan a una resolución espacial de 84x128, mientras que para T2W se redimensionan a una resolución espacial de 320x320.

Por otro lado, dado que el conjunto de datos es muy pequeño y presenta un desbalance entre las dos clases, se realizó sobremuestreo (*oversampling*) o aumento de datos (*data augmentation*), con el fin de aumentar en número de ejemplos en cada modalidad por clase. Este proceso básicamente consistió en generar nuevas imágenes al girar y voltear los parches originales, hasta obtener siete nuevos ejemplos, tal como se ilustra en la Figura 3.

Diseños de las arquitecturas de las CNN

En el diseño de las arquitecturas de las CNN, se plantearon cuatro arquitecturas básicas basadas en la arquitectura original de *LeNet-5* como lo hicieron los trabajos pioneros de visión por computador para la AlexNet (Krizhevsky, n.d),y VGG16 (Simonyan, 2018), pero con una tres capas de convolución y *pooling*, y una única última capa densa totalmente conectada, variando el número de neuronas por capa de cada arquitectura para su comparación (Figura 4). Esto, con el objetivo de analizar la complejidad y el desempeño de las *CNN*, con cada modalidad de manera independiente (unimodales) y con los dos conjuntos de datos bimodales proporcionando una capa adicional de la clásica LeNet-5 para el aprendizaje de las características visuales con una capa más de convolución y *pooling*. Sin embargo, también se adaptó una arquitectura del estado del arte probada con el conjunto datos proporcionado por el *PROSTATEx Challenge*, de acuerdo con el diseño de la arquitectura de la red neuronal convolucional de Gutiérrez *et al.*, (Gutiérrez, 2020) basada directamente en la arquitectura de LeNet-5, i.e. dos capas de convolución y pooling, seguida de tres capas densas totalmente conectadas. En la Figura 5 se puede observar el diseño de la arquitectura y la Tabla 6 detalla las capas que la componen.

Entrenamiento de las arquitecturas con el conjunto de datos

Para el desarrollo del entrenamiento se aplicó la técnica de validación cruzada de K iteraciones (*K-fold cross validation* en inglés) con K=5, correspondiente al número de subconjuntos en los que se divide el conjunto de datos de forma proporcional o estratificada de acuerdo con el número de muestras por clase. El objetivo de aplicar esta técnica es que permite ver si existe independencia entre los datos y evaluar la precisión del modelo (Shultz *et al.*, 2011).

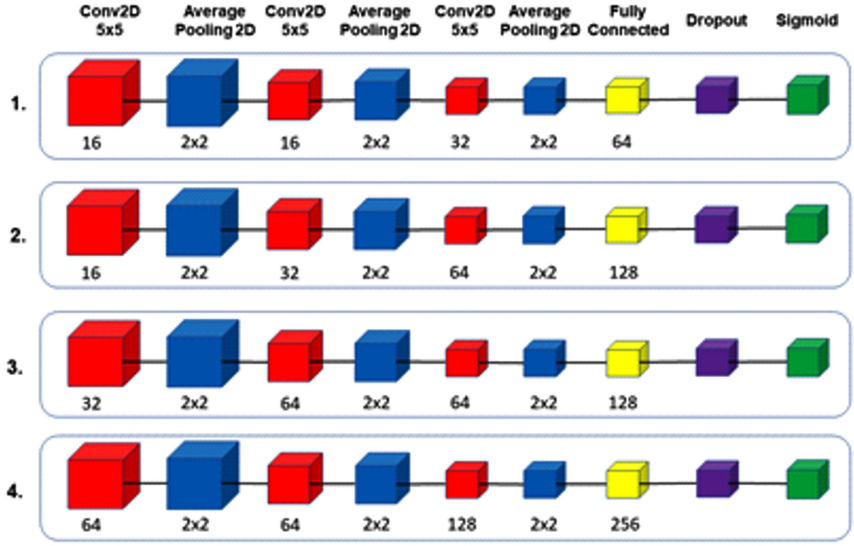

Figura 4. Cuatro arquitecturas CNN utilizadas en la fase inicial.

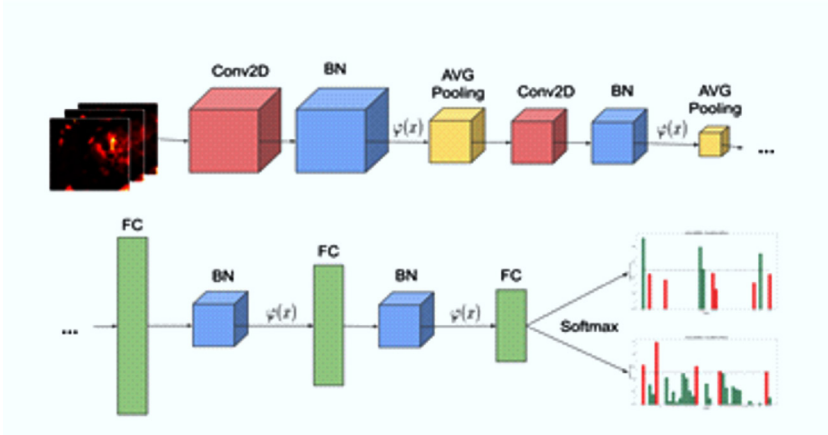

Figura 5. Diseño de la arquitectura desarrollada por Gutiérrez et. al, inspirada en la LeNet (Gutiérrez, 2019) usada como base para la arquitectura 5.

Un subconjunto (*fold)* es una división que contiene un parte de los datos, cada uno está compuesto por dos clases (alto y bajo), y cada clase contiene 120 regiones cuadradas (parches) con las lesiones, por lo que en total cada subconjunto se compone 240 parches (ver Figura 6), Esto aplica para el conjunto de datos unimodales y los dos conjuntos bimodales.

Además, en la fase del entrenamiento, la combinación de las modalidades *Ktrans-T2W* y *Ktrans-ADC* se realiza a la entrada del modelo de la *CNN* al ingresar como una única imagen de dos canales con cada modalidad por canal, a lo cual se conoce *early fusion*, y se realizó con el objetivo de comparar las estrategias de fusión con los resultados del trabajo de Gutiérrez *et al*., (Gutiérrez, 2020), donde hicieron *late fusion* al combinar de manera ponderada la contribución de las salidas del clasificador *CNN* entrenado para cada modalidad independiente en una única salida, donde pesaba seis veces más la modalidad Ktrans y las demás (T2 transaxial, T2 sagital, ADC) con pesos iguales.

Los parámetros explorados para la arquitectura de las CNN con sus respectivos valores de forma aleatoria de acuerdo con trabajos previos fueron: *dropout* = [0.2,0.3,0.4,0.5], *learning rate* = [0.001,0.0009,0.0005,0.0001], *learning rate decay* = [0.0001,0.01,0.001,0.0001], y para el entrenamiento se utilizó el optimizador Adam (*Adaptive moment estimation*).

Medidas de desempeño

Las medidas de desempeño permiten tener una métrica para evaluar los resultados obtenidos al implementar un modelo en un conjunto de datos de prueba. Las métricas más utilizadas se construye a partir de la matriz de confusión, la cual describe los siguientes datos: los verdadero-positivos (TP) son aquellas muestras que pertenece a la clase positiva y que son predichos de manera positiva, los verdaderos-negativos (TN) son aquellas muestras que pertenece a la clase negativas y que son predichos de manera negativa, los falsos-positivos (FP) son aquellas muestras que pertenece a la clase negativa pero son predichos de manera positiva, y los falsos-negativos (FN) son aquellas muestras que pertenece a la clase positiva pero son predichos como de la clase negativa. En la Tabla 6 se describen las medidas de desempeño utilizadas (Sunasra, 2017).

Tabla 6. Descripción de capas que componen la arquitectura CNN

Capas	Número de Capas
Conv2D 3x3 filter	2
Batch Normalization	4
Average Pooling	2
Fully connected	3
ReLu	4
Softmax	1

Es importante mencionar que el conjunto de datos tomado del *ProstateX Challenge* establece como métrica para evaluar el desempeño de los modelos y de los resultados obtenidos el *AUC* (área bajo la curva ROC), la cual muestra la probabilidad que tiene un clasificador

para que de manera aleatoria clasifique un ejemplo positivo como realmente positivo frente a un ejemplo que es negativo y se ha clasificado como positivo, entre más cerca esté el valor del *AUC* al uno indica que tuvo más predicciones correctas (Google Developers, 2020).

Resultados

En los resultados obtenidos por validación cruzada (K=5), de los diferentes conjuntos de datos que se conformaron para cada modalidad independiente (unimodal) y en la unión de las modalidades *Ktrans-T2W* y *Ktrans- ADC* (bimodales), se calcularon las medidas de desempeño para hacer una comparación de qué modalidad independiente (unimodal) o combinación de modalidades (bimodales) obtuvo mejor rendimiento en el entrenamiento. Además, se midió el desempeño de las arquitecturas con el uso de la Curva ROC, la cual es una representación gráfica de la sensibilidad frente a la especificidad para un clasificador binario según se varía el umbral de discriminación, permite observar los puntos de corte donde la sensibilidad y la especificidad es más alta, y además tiene la capacidad de discriminar las pruebas diagnósticas con el fin de diferenciar casos con lesiones de riesgo alto versus lesiones de riesgo bajo (Cerda y Cifuentes, 2012).

En la fase inicial se evaluaron todos los conjuntos de datos unimodal y bimodal. Además, se hicieron varios entrenamientos utilizando diferentes modelos, con variaciones de parámetros en las 4 arquitecturas propuestas. En total se trabajó con 64 modelos diferentes, con las respectivas variaciones en los parámetros de las *CNN*, los cuales fueron *dropout, learning rate, learning rate decay*, con esto se permite escoger los mejores desempeños con la configuración más eficiente para cada conjunto de datos.

- ***Arquitectura 1 o CNN 5L-16n-64n***

En la Tabla 7, se describe los valores de los promedios de las medidas de desempeño de cada uno de los conjuntos de datos unimodales y bimodales. Se puede observar que la modalidad *Ktrans* fue la que obtuvo mejor rendimiento con un *AUC* promedio de 0.70 ± 0133 con la implementación de la arquitectura *CNN 5L-16n-64n*. En la Figura 7 se puede observar la curva ROC y el valor del *AUC* en cada uno de los subconjuntos (*folds),* el fold 3 presenta el mejor desempeño con un AUC de 0.88 mientras el fold 5 el peor con un AUC 0.55 y estos los conforman el conjunto de datos de la modalidad *Ktrans*.

- ***Arquitectura 2 o CNN 5L-16n-128n***

En la Tabla 8, se describe los valores de los promedios de las medidas de desempeño de cada uno de los conjuntos de datos unimodales y bimodales. Se puede observar que la modalidad *Ktrans* fue la que obtuvo mejor rendimiento con un *AUC* promedio de 0.70 ± 0.129 con la implementación de la arquitectura *CNN 5L-16n-128n*. En la Figura 8 se puede observar la curva ROC y el valor del *AUC* en cada uno de los subconjuntos (*folds),* el fold 3 presenta el mejor desempeño con un AUC de 0.88 mientras el fold 5 el epor con un AUC 0.55 y estos los conforman el conjunto de datos de la modalidad *Ktrans*.

- ***Arquitectura 3 o CNN 5L-32n-128n***

En la Tabla 9, se describe los valores de los promedios de las medidas de desempeño de cada uno de los conjuntos de datos unimodales y bimodales. Se puede observar que la modalidad *Ktrans* fue la que obtuvo mejor rendimiento con un *AUC* promedio de 0.71 ± 0.127 con la implementación de la arquitectura *CNN 5L-32n-128n*. En la Figura 9 se puede observar la curva ROC y el valor del *AUC* en cada uno de los subconjuntos (*folds),* el fold 3 presenta el mejor desempeño con un AUC de 0.89 mientras el fold 5 el peor con un AUC 0.56 y estos los conforman el conjunto de datos de la modalidad *Ktrans*.

- ***Arquitectura 4 o CNN 5L-64n-256n***

En la Tabla 10, se describe los valores de los promedios de las medidas de desempeño de cada uno de los conjuntos de datos unimodales y bimodales. Se puede observar que la modalidad *Ktrans* fue la que obtuvo mejor rendimiento con un *AUC* promedio de 0.70 ± 0.135 con la implementación de la arquitectura *CNN 5L-64n-256n*. En la Figura 10 se puede observar la curva ROC y el valor del *AUC* en cada uno de los subconjuntos (*folds),* el fold 3 presenta el mejor desempeño con un AUC de 0.89 mientras el fold 5 el peor con un AUC 0.54 y estos los conforman el conjunto de datos de la modalidad *Ktrans*.

- ***Arquitectura 5***

Para la fase posterior se evaluaron los conjuntos de datos unimodales (*Ktrans, ADC, T2W) y bimodales (Ktrans-ADC, Ktrans-T2W*) con la arquitectura descrita en la Figura 5, con una modificación en la última capa reem-

Tabla 7. Descripción de las medidas de desempeño.

Medida	Descripción	Ecuación
Exactitud (*Accuracy*)	La exactitud en los problemas de clasificación permite encontrar la proporción de predicciones correctas, que fueron realizadas por el modelo sobre todas las predicciones hechas.	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensibilidad (*Sensitivity*)	La sensibilidad nos da la proporción de los casos positivos sobre todos los casos predichos por el modelo como positivos.	$\frac{TP}{TP + FN}$
Especificidad (*Specificity*)	La especificidad nos da la proporción de los casos negativos sobre todos los casos predichos por el modelo como negativos.	$\frac{TN}{TN + FP}$
Precisión (*Precision*)	La precisión nos da la proporción de los casos que pertenece a los casos positivos sobre el total de las predicciones que se hicieron de manera correcta.	$\frac{TP}{TP + FP}$
F1 Score	El F1 Score se define como una media armónica entre la precisión y la sensibilidad.	$\frac{2TP}{2TP + FP + FN}$
Área bajo la curva (*Area Under Curve*)	El *AUC* se puede entender como una medida de separabilidad, muestra la capacidad que tiene un modelo de distinguir las diferentes clases, cuanto más alto es el valor del *AUC* mejor es el modelo para predecir las clases positivas como positivas y las negativas como negativas (Narkhede, 2018).	0.9 - 1.0 = Excelente 0.8 - 0.9 = Bueno 0.7 - 0.8 = Regular 0.6 - 0.7 = Malo 0.5 - 0.6 = Deficiente

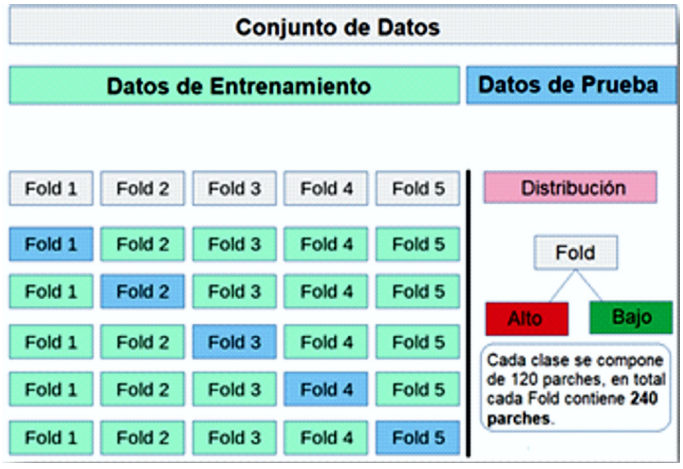

Figura 6. Representación de la aplicación de la técnica K-fold Cross Validation (Validación Cruzada con K iteraciones) con K=5.

plazando la capa *softmax* por una *sigmoide*, debido a que la capa *softmax* está diseñada para clasificaciones multiclase y la sigmoide se usa para clasificaciones binarias Posteriormente, en el entrenamiento se calcula las medidas desempeño en cada uno de los subconjuntos (*folds*) para posteriormente obtener el promedio de cada una de las medidas desempeño, que se encuentran descritas en la Tabla 11. Se puede observar que la modalidad *Ktrans-T2W* fue la que obtuvo mejor rendimiento con un *AUC* promedio de 0.72 ± 0.058, con la implementación de la arquitectura Gutiérrez et al. En la Figura 11 se puede observar la curva ROC y el valor del *AUC* en cada uno de los subconjuntos (*folds*), el fold 2 presenta el mejor desempeño con

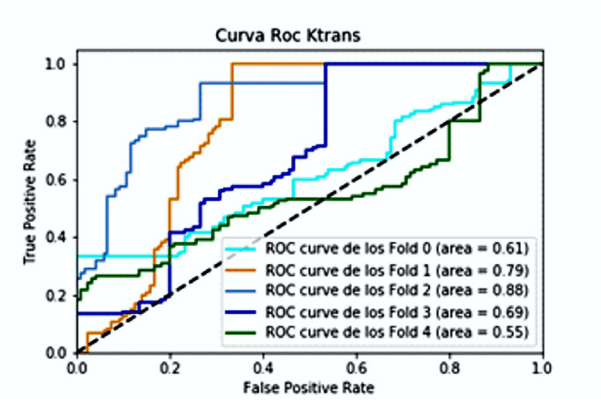

Figura 7. Curva ROC del conjunto de datos Ktrans en la arquitectura 1.

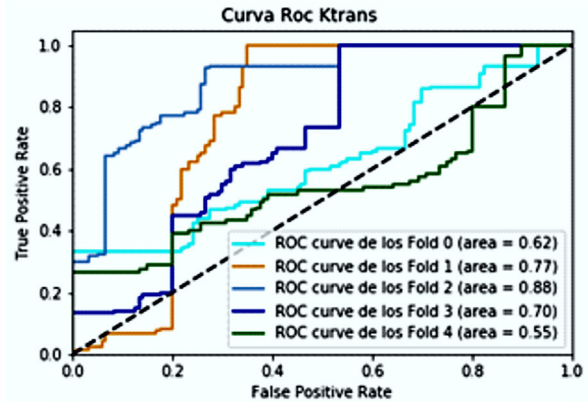

Figura 8. Curva ROC del conjunto de datos Ktrans en la arquitectura 2.

Tabla 8. Descripción de los promedios de las medidas de desempeño para cada conjunto de datos con la arquitectura 1.

Conjuntos de Datos	Exactitud	Sensibilidad	Especificidad	Precisión	F1 Score	AUC
Ktrans	**0.664 ±0.120**	0.776 ±0.175	0.551 ±0.081	**0.627 ±0.086**	**0.691 ±0.123**	**0.70 ±0.133**
ADC	0.5 ±0.0	0.6 ±0.489	0.4 ±0.489	0.3 ±0.245	0.4 ±0.326	0.5 ±0.0
T2W	0.514 ±0.028	0.778 ±0.391	0.25 ±0.387	0.408 ±0.204	0.535 ±0.267	0.518 ±0.040
Ktrans-ADC	0.595 ±0.084	**0.836 ±0.250**	0.355 ±0.386	0.595 ±0.084	0.657 ±0.103	0.582 ±0.131
Ktrans-T2W	0.591 ±0.093	0.626 ±0.339	**0.556 ±0.336**	0.486 ±0.255	0.536 ±0.272	0.612 ±0.128

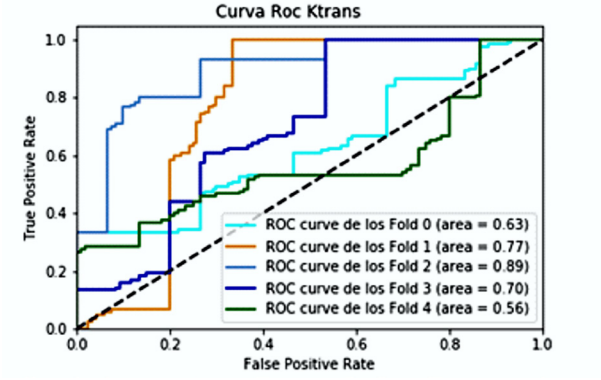

Figura 9. Curva ROC del conjunto de datos Ktrans en la arquitectura 3.

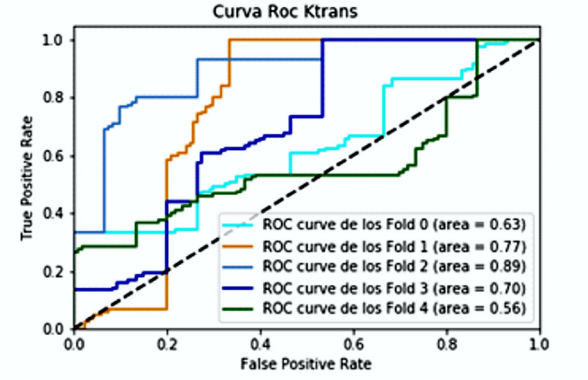

Figura 10. Curva ROC del conjunto de datos Ktrans en la arquitectura 4

Tabla 9. Descripción de los promedios de las medidas de desempeño para cada conjunto de datos con la arquitectura 2.

Conjuntos de Datos	Exactitud	Sensibilidad	Especificidad	Precisión	F1 Score	AUC
Ktrans	**0.668 ±0.129**	0.755 ±0.180	**0.581 ±0.097**	**0.636 ±0.100**	**0.688 ±0.133**	**0.70 ±0.129**
ADC	0.5 ±0.0	0.6 ±0.489	0.4 ±0.489	0.3 ±0.245	0.4 ±0.326	0.5 ±0.0
T2W	0.523 ±0.047	0.566 0.466	0.480 ±0.448	0.316 ±0.260	0.403 ±0.329	0.56 ±0.092
Ktrans-ADC	0.564 ±0.085	**0.765 0.385**	0.365 0.369	0.443 0.228	0.558 0.281	0.54 0.047
Ktrans-T2W	0.602 ±0.103	0.648 ±0.358	0.556 ±0.336	0.492 ±0.258	0.546 ±0.280	0.50 ±0.0

Tabla 10. Descripción de los promedios de las medidas de desempeño para cada conjunto de datos con la arquitectura 3.

Conjuntos de Datos	Exactitud	Sensibilidad	Especificidad	Precisión	F1 Score	AUC
Ktrans	**0,665 ±0.140**	**0,765 ±0.177**	0,565 ±0.113	**0,631 ±0.110**	**0,690 ±0.137**	**0,71 ±0.127**
ADC	0.5 ±0.0	0.6 0.489	0.4 ±0.489	0.3 ±0.245	0.4 ±0.326	0.5 ±0.0
T2W	0.505 ±0.011	0.75 ±0.387	0,265 ±0.388	0.404 ±0.202	0.521 ±0.261	0.45 ±0.073
Ktrans-ADC	0.542 ±0.084	0.526 ±0.450	0.558 ±0.462	0.350 ±0.300	0.404 ±0.330	0.50 ±0.040
Ktrans-T2W	0.576 ±0.102	0.54 ±0.446	0.613 ±0.373	0.358 ±0.230	0.426 ±0.351	0.50 ±0.0

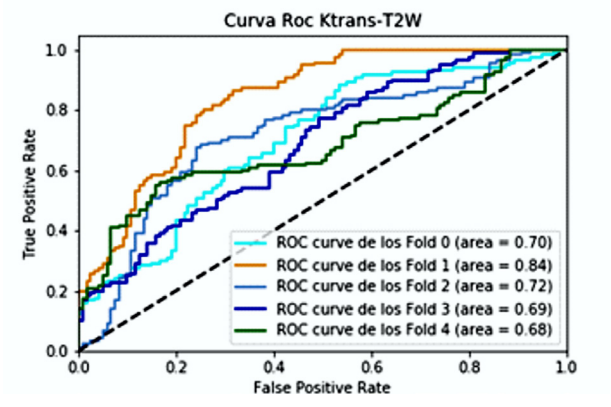

Figura 11. Curva ROC del conjunto de datos Ktrans-T2W con la arquitectura 5.

un AUC de 0.84 mientras el fold 5 el peor con un AUC 0.68 y estos los conforman en conjunto de datos de la modalidad *Ktrans-T2W*.

En general, se puede evidenciar que el en las arquitecturas de 1 al 4 el fold 4 fue el que presentó el rendimiento más bajo y el fold 3 el más alto pero en la arquitectura 5, el fold 2 presentó el mejor rendimiento y todos lo valores estuvieron más agrupados, puede significar que obtuvieron resultados más robustos.

Costo computacional

En la tabla 13 se muestran los tiempos de ejecución en los entrenamientos de los modelos *CNN*, para las cinco arquitecturas, entrenados tanto con datos uni-

Tabla 11. Descripción de los promedios de las medidas de desempeño para cada conjunto de datos con la arquitectura 4.

Conjuntos de Datos	Exactitud	Sensibilidad	Especificidad	Precisión	F1 Score	AUC
Ktrans	**0.667 ±0.125**	**0.760 ±0.181**	0.575 ±0.081	**0.633 ±0.093**	**0.688 ±0.130**	**0.70 ±0.135**
ADC	0.5 ±0.0	0.4 ±0.489	0.6 ±0.489	0.2 ±0.245	0.266 ±0.326	0.5 ±0.0
T2W	0.513 ±0.026	0.386 ±0.474	0.64 ±0.490	0,207 ±0.254	0.270 ±0.330	0.544 ±0.083
Ktrans-ADC	0.575 ±0.091	0.50 ±0.437	0.65 ±0.388	0.377 ±0.319	0.411 ±0.337	0.50 ±0.0
Ktrans-T2W	0.583 ±0.082	0.375 ±0.328	**0.791 ±0.190**	0,388 ±0.320	0.375 ±0.313	0.50 ±0.0

Tabla 12. Descripción de los promedios de las medidas de desempeño para cada conjunto de datos basado en la arquitectura 5.

Conjuntos de Datos	Exactitud	Sensibilidad	Especificidad	Precisión	F1 Score	AUC
Ktrans	0.5875 ±0.042	0.5867 ±0.119	0.5883 ±0.0742	0.586 ±0.0414	0.5826 ±0.0747	0.672 ±0.027
ADC	0.531 ±0.096	0.583 ±0.130	0.478 ±0.100	0.526 ±0.090	0,552 ±0,103	0.536 ±0.111
T2W	0.559 ±0.049	0.560 ±0.090	0.558 ±0.075	0.559 ±0.048	0,557 ±0,065	0.583 ±0.049
Ktrans-ADC	0.578 ±0.067	0.642 ±0.122	0.515 ±0.094	0.569 ±0.064	0.600 ±0.081	0.62 ±0.068
Ktrans-T2W	**0.636 ±0.052**	**0.613 ±0.084**	**0.658 ±0.141**	**0.657 ±0.096**	**0.627 ±0.043**	**0.72 ±0.058**

modales como bimodales. Se evidencia tiempos ligeramente más cortos cuando se tienen arquitecturas más sencillas con menos unidades (neuronas) por capa. Sin embargo, todos los tiempos de ejecución en los entrenamientos de los modelos *CNN* presentados en la tabla son inferiores a un minuto gracias a que este proceso se realizó usando una tarjeta de procesamiento gráfico para acelerarlo, en este caso se usó una una NVIDIA GTX Titan X con 12GB de memoria y 3072 CUDA cores.

Discusión

Los resultados obtenidos presentan un desempeño comparable con los alcanzados por (Gutiérrez, 2020) y por encima de los reportados por (Gutiérrez, 2019) con los que es directamente comparable, sin embargo, tiene la ventaja de obtener un desempeño similar usando solo dos modalidades en lugar de cuatro, lo cual aligera el costo de uso de memoria y procesamiento computacional en la práctica de la *CNN* obtenida, tanto para entrenar como para su uso en predicción. Por otro lado, nuestra evaluación alcanzó resultados más bajos al compararlos con los trabajos de Saifeng *et al.*, (Saifeng, 2017), Alireza *et al.*, (Alireza, 2017) y Jarrel *et al.*, (Jarrel, 2017) que usaban CNN para la tarea de clasificación donde se enfoca en clasificar las lesiones en dos tipos de clases, indolentes (riesgo bajo) y clínicamente significativo (riesgo alto) en el mismo conjunto de datos de PROSTATEx. En cuanto al análisis de porqué se usan las *CNN* u otros métodos de clasificación, existen algunas puntos esenciales que pue-

Tabla 13. Reporte de los tiempos de ejecución de los modelos en las diferentes arquitecturas.

CNN	Ktrans	ADC	T2W	Ktrans-ADC	Ktrans-T2W
Arquitectura 1	54,32 seg	54,18 seg	54,47 seg	54,03 seg	54,22 seg
Arquitectura 2	56,13 seg	54,58 seg	53,32 seg	57,74 seg	56,27 seg
Arquitectura 3	53,04 seg	52,79 seg	53,33 seg	55,31 seg	56,17 seg
Arquitectura 4	57,02 seg	56,63 seg	55,47 seg	56,02 seg	56,54 seg
Arquitectura 5	57,32 seg	56,53 seg	56,27 seg	55,28 seg	56,31 seg

den ser los siguientes, las *CNN* en comparación a otros métodos clasificación han sido diseñadas para trabajar con imágenes y son capaces de aprender las características representativas de una imagen por sí solas, en cambio, otros métodos, por ejemplo, los basados en *SVM*, está diseñados para ser clasificadores genéricos y su rendimiento está sujeto a la cantidad de características específicamente diseñadas y seleccionadas para el contexto que se usan para clasificar y maximizar el margen entre las diferentes clases. Sin embargo, es importante mencionar que las *CNN* pueden tener un costo computacional mayor si se pretende hacer una red con muchas capas, si se usan más modalidades de imágenes de entrada y sí los valores de los parámetros en dichas capas no son los adecuados. Las regiones cuadradas de las lesiones, a partir de las imágenes médicas de resonancia magnética, tienen una particularidad asociada a que por el tipo de imagen diagnóstica, resolución y área anatómica, no tienen formas o bordes bien detallados. Por lo anterior, el uso de las cuatros primeras arquitecturas que cuentan con un número mayor de neuronas y capas puede no serla mejor configuración ya que se puede sobreentrenar el modelo, o ser más complejo de lo requerido para esta tarea, perdiendo o diluyendo las características relevantes en la última capa de convolución y *pooling*. Lo contrario sucede con la arquitectura 5, que se había probado previamente en el mismo conjunto de datos, pero no en el escenario de fusión de dos modalidades (bimodales), con la cual se obtiene mejora en el desempeño de la modalidad unimodal *Ktrans*. Además, la arquitectura 5 cuenta con tres capas densas completamente conectadas en contraste con las arquitecturas de la 1 a la 4, lo cual resalta la importancia de estas capas para la selección de características de alto nivel y la reducción de dimensionalidad de su representación final. Esto evidencia que más que tener más capas y más neuronas para la representación en las imágenes de las capas de convolución y *pooling*, tiene mayor impacto las capas finales densas completamente conectadas.

Con respecto a los trabajos presentados en la Tabla 1, se puede observar que los resultados obtenidos son significativamente altos, sin embargo es importante precisar que en esos trabajos las CNN usaron el conjunto de datos en sus dimensiones 3D y no planos 2D como en los trabajos de (Gutiérrez, 2020) y este, permitiendo aprovechar más información y utilizando técnicas de segmentación para diferenciar las regiones de interés, aunque en términos computacionales son más exigentes en comparación al proceso realizado en este trabajo.

Cabe mencionar que con una cantidad de datos de entrenamiento mayor se podría llegar a obtener un mejor rendimiento. Los resultados obtenidos son equivalentes al desempeño alcanzado por (Gutiérrez *et al*., 2020). Sin embargo, en comparación, la estrategia *late fusion* usada en el trabajo de Gutiérrez et al. y la estrategia *early fusion* usada en este trabajo, no se logran determinar un desempeño mejor, y por tanto no representa una diferencia muy significativa, aunque el enfoque nuestro solo requiere dos modalidades teniendo un método más compacto y ligero para entrenamiento posterior con más datos o su uso en predicción. En cuanto a la naturaleza 3D de los datos, con los resultados que se muestran en la Tabla 1, demuestra que se puede aprovechar para mejorar el desempeño de los modelos, esto permitiría que se aprendan mejor las características y relaciones entre las modalidades y alcanzar no solo desempeños similares sino incluso superiores a los obtenidos en este trabajo, lo cual motiva el trabajo futuro y experimentación adicional en esta línea de investigación. El desafío estaría en aprovechar esta información sin que eso conlleve a una mayor carga computacional de datos a procesar o en modelos más costosos computacionalmente.

Con respecto a valores descritos en las Tablas 8 - 11, se observa que en la tabla 8, la modalidad que mejor tuvo resultado fue la Ktrans, donde alcanzó un AUC promedio de 0.70 (±0.133) , en la tabla 9 la modalidad que mejor tuvo resultado fue la Ktrans, con un AUC promedio de 0.70 (± 0.129), en la tabla 10, la modalidad que tuvo mejor rendimiento fue la Ktrans con un AUC promedio de 0.71 (±0.127), en la tabla 11, la modalidad con mejor desempeño es la Ktrans con un AUC promedio de 0.70 (± 0.135), en cuanto a las tabla 12, las modalidades que alcanzaron un mejor desempeño fue la combinación de la Ktrans-T2W alcanzando un AUC promedio 0.72 (±0.058). Es de destacar que las modalidades ADC y T2W por sí solas no presentan un buen rendimiento en la clasificación en cada una de las arquitecturas y se puede evidenciar en las medidas de desempeño utilizadas, como, por ejemplo, la exactitud dónde describe que prácticamente las predicciones en los modelos en dichas modalidades rondan entre 50% - 55% de clasificación entre las dos clases, casos contrario, con la modalidad Ktrans.

En cuanto a los resultados obtenidos, se evidencia en las Tablas 8-11 que la modalidad *Ktrans* presenta un mejor desempeño para los modelos al momento de clasificar tal como ocurre y se está incorporando y priorizando en la práctica clínica (Vos *et al*., 2013), esto debido a que esta modalidad presenta un mejor contraste en el tejido de interés y en las lesiones cancerosas, que puede facilitar el entrenamiento de los modelos para el aprendizaje de las representaciones en cada clase. Sin embargo, en la Tabla 12 se puede observar que la combinación de las modalidades *Ktrans y T2W* obtuvo un valor mayor en *AUC*, pero la diferencia con solo la modalidad *Ktrans* es de aproximadamente 5%, o cual significa que *T2W* complementa a *Ktrans* para mejorar el desempeño de *Ktrans* de forma independiente, y refuerza el hecho que la modalidad *Ktrans* es la que proporciona la mayor información a la CNN para que el modelo aprenda y clasifique entre los dos grupos de lesiones de cáncer de próstata, de bajo riesgo y alto riesgo.

Conclusiones

La aplicación de estrategias de visión por computador del estado del arte en el área imagen médica han demostrado resultados prometedores para la detección y el diagnóstico de enfermedades. En el caso particular de este trabajo, fueron utilizadas para la diferenciación del grado de agresividad de lesiones de cáncer de próstata. Sin embargo, es necesario que se disponga de la información necesaria para implementar y usar las redes neuronales convolucionales, principalmente grandes volúmenes de datos con las anotaciones de la ubicación de las lesiones. Estos datos médicos son escasos, ya que se cuenta con pocos conjuntos de datos públicos disponibles y de los que existen no todos se encuentran debidamente anotados. Con el uso de métodos de aprendizaje profundo (*Deep Learning*), las *CNNs* permiten obtener resultados o predicciones eficientes y cuantificables en el diagnóstico del cáncer de próstata, superando obstáculos como es caso el de no contar con una gran cantidad de datos. De esta forma, es posible proporcionar un apoyo a los radiólogos para el diagnóstico temprano y estimación apropiada del grado de agresividad del cáncer de próstata, contribuyendo directamente en la calidad de vida de los pacientes.

Es fundamental destacar que para el uso de las redes neuronales convolucionales es conveniente contar con una arquitectura sólida y robusta que permita explotar con la mayor eficiencia los datos que se tienen disponibles. En la comparación de las diferentes arquitecturas con respecto a la arquitectura que se encuentra bien estructurada y validada, son notables los aspectos de rendimiento en el entrenamiento de los modelos y la clasificación de los datos. En el área de imágenes médicas, en especial en patrones similares a texturas, no siempre la arquitectura con más neuronas y más capas (más profunda) es la que obtiene mejores resultados. Además, este tipo de imágenes médicas presenta un gran reto a la hora de clasificar una lesión en un estadio de severidad, dado que no proporciona mucha información visual de apoyo para este tipo de imágenes médicas, a excepción de modalidades como *Ktrans* que tienen mejor especificidad para los radiólogos, por lo tanto, siempre requiere la revisión posterior para los radiólogos en el contexto de radiómica, y aprovechar estos métodos como apoyo al diagnóstico de cáncer en estadios tempranos.

Agradecimientos

Los autores agradecen al grupo de investigación CIM@LAB de la Universidad Nacional de Colombia, al Grupo de Investigación GITECX y al Semillero de Investigación AdaLab de la Universidad de los Llanos, por el apoyo en la realización de este proyecto. Este trabajo fue parcialmente soportado gracias al proyecto C09-F02-011-2019 de la DGI/Unillanos y el proyecto BPIN 2019000100060 del FCTeI del SGR.

Referencias

Lafuente-Martínez J, Luis Y, Moreno H. (2016). GENERALIDADES Y Conceptos Básicos de Resonancia Magnética (RM) Técnica de la Imagen por Resonancia Magnética. Retrieved from http://www.serme.es/wp-content/uploads/2016/05/capitulo1p.pdf

Global Cancer Observatory. (2018). Age standardized (World) incidence rates, prostate, all ages. https://doi.org/10.6

American Cancer Society. (2017). Pruebas para detectar el cáncer de próstata. Retrieved September 27, 2019, from https://www.cancer.org/es/cancer/cancer-de-prostata/deteccion-diagnostico-clasificacion-por-etapas/como-se-diagnostica.html

Barentsz JO, Weinreb JC, Verma S, Thoeny HC, Tempany CM, Shtern F, Choyke PL. Synopsis of the PI-RADS v2 Guidelines for Multiparametric Prostate Magnetic Resonance Imaging and Recommendations for. Use. *European Urology*. 2016;69(1):41-49. https://doi.org/10.1016/j.eururo.2015.08.038

Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, Deasy JO. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 2015;112(46):E6265-73. https://doi.org/10.1073/pnas.1505935112

Wibmer A, Hricak H, Gondo T, Matsumoto K, Veeraraghavan H, Fehr D, Vargas HA. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. European Radiology, 2015;25(10):2840-2850. https://doi.org/10.1007/s00330-015-3701-8

Tiwari P, Kurhanewicz J, Madabhushi A. Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. *Medical Image Analysis*. 2013;17(2):219-235. https://doi.org/10.1016/J.MEDIA.2012.10.004

Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng K-T. (Tim), Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. *Physics in Medicine & Biology*. 2017;62(16): 6497-6514. https://doi.org/10.1088/1361-6560/aa7731

Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? In on Medical Imaging (Vol. 35).

Karen Simonyan, Andrew Zisserman. "very deep convolutional networks for large-scale image recognition Karen." *American Journal of Health-System Pharmacy*. 2018;75(6):398-406.

Krizhevsky Alex, Ilya Sutskever, Geoffrey E. Hinton. n.d. ImageNet Classification with Deep Convolutional Neural Networks.

Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift Fur Medizinische Physik, 2019;29:102-127. https://doi.org/10.1016/j.zemedi.2018.11.002.

Gutiérrez Y, Arevalo J, Martínez F. "A Ktrans deep characterization to measure clinical significance regions on prostate cancer," Proc. SPIE 11330, 15th International Symposium on Medical Information Processing and Analysis, 113300C (3 January 2020); https://doi.org/10.1117/12.2542606

Gutiérrez Y, Garzón G, Martínez F. (2019, April 1). Towards clinical significance prediction using ktrans evidences in prostate cancer. 2019 22nd Symposium on Image, Signal Processing and Artificial Vision, STSIVA 2019 - Conference Proceedings. https://doi.org/10.1109/STSIVA.2019.8730282

Shultz, Thomas R, Scott E. Fahlman, Susan Craw, Periklis Andritsos, Panayiotis Tsaparas, Ricardo Silva, Chris Drummond, Charles X. Ling, Victor S. Sheng, Chris Drummond, Pier Luca Lanzi, João Gama, R. Paul Wiegand, Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Jun He, Sanjay Jain, Frank Stephan, Sanjay Jain, Frank Stephan, Claude Sammut, Michael Harries, Claude Sammut, Kai Ming Ting, Bernhard Pfahringer, John Case, Sanjay Jain, Kiri L. Wagstaff, Siegfried Nijssen, Anthony Wirth, Charles X. Ling, Victor S. Sheng, Xinhua Zhang, Claude Sammut, Nicola Cancedda, Jean-Michel Renders, Pietro Michelucci, Daniel Oblinger, Eamonn Keogh, and Abdullah Mueen. 2011. "Cross-Validation." Pp. 249–249 in Encyclopedia of Machine Learning.

Sunasra, Mohammed. 2017. "Performance Metrics for Classification Problems in Machine Learning." Retrieved March 9, 2020 (https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-ib085d432082b).

GoogleDevelopers. 2020. "Clasificación: ROC y AUC | Curso Intensivo de Aprendizaje Automático." 2020. Retrieved June 7, 2021 (https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc).

Cerda, Jaime, and Lorena Cifuentes. 2012. "Uso de Curvas ROC En Investigación Clínica. Aspectos Teórico-Prácticos." Revista Chilena de Infectologia 29(2):138–41. doi: 10.4067/S0716-10182012000200003.

Narkhede, Sarang. 2018. "Understanding AUC - ROC Curve." Retrieved March 9, 2020 (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5).

Liu, Saifeng, Huaixiu Zheng, Yesu Feng, and Wei Li. 2017. "Prostate Cancer Diagnosis Using Deep Learning with 3D Multiparametric MRI." P. 1013428 in Medical Imaging 2017: Computer-Aided Diagnosis. Vol. 10134. SPIE.

Liu, Yongkai, Kyunghyun Sung, Guang Yang, Sohrab Afshari Mirak, Melina Hosseiny, Afshin Azadikhah, Xinran Zhong, Robert E. Reiter, Yeejin Lee, and Steven S. Raman. 2019. "Automatic Prostate Zonal Segmentation Using Fully Convolutional Network with Feature Pyramid Attention." IEEE Access 7:163626–32. doi: 10.1109/ACCESS.2019.2952534.

Mehrtash, Alireza, Alireza Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M. Tempany, William M. Wells, Tina Kapur, Parvin Mousavi, Purang Abolmaesumi, and Andriy Fedorov. 2017. "Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks." P. 101342A in Me-

dical Imaging 2017: Computer-Aided Diagnosis. Vol. 10134. SPIE.

Seah, Jarrel C. Y., Jennifer S. N. Tang, and Andy Kitchen. 2017. "Detection of Prostate Cancer on Multiparametric MRI." P. 1013429 in Medical Imaging 2017: Computer-Aided Diagnosis. Vol. 10134. SPIE.

Vos, E. K., Litjens, G. J. S., Kobus, T., Hambrock, T., Kaa, C. A. H. Van De, Barentsz, J. O., Huisman, H. J., & Scheenen, T. W. J. (2013). Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 T. European Urology, 64(3), 448–455. https://doi.org/10.1016/j.eururo.2013.05.045