Universidad Internacional de La Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

Comparativa de Técnicas de Preprocesamiento y Entrenamiento para Detección de Hemorragias Intracraneales

Trabajo Fin de Máster

Presentado por: Arroyo Rubio, Fernando

Director/a: Mancera Valetts, Laura

Ciudad: Madrid
Fecha: 22/02/2020

Contenido

Índice de Tablas

Índice de Ilustraciones

Resumen

En las últimas dos décadas, el mundo de la medicina está teniendo un notable interés por el uso del procesamiento de imágenes con Inteligencia Artificial para la detección de numerosas enfermedades y lesiones que tienen lugar en el ser humano. Específicamente, las técnicas de Machine Learning y Deep Learning han dejado ver sus bondades para ayudar a reducir el error humano y apoyar a que los diagnósticos sean más acertados de los que se obtienen a través de métodos tradicionales tratando de encontrar patrones ocultos dentro de las imágenes a procesar. En este Trabajo de Fin de Máster, se aplican diferentes técnicas de preprocesamiento de imágenes y algoritmos de entrenamiento para la detección e identificación de distintos tipos de Hemorragias Intracraneales (ICH) a través de tomografías computarizadas (TC). Todo esto se implementa en un escenario experimental con el objetivo de conocer qué técnica y algoritmo ofrece un mejor resultado. Los resultados demuestran que el uso de redes neuronales convolucionales, técnicas de preprocesado como Windowing y el empleo de métodos de balanceo de datos permiten elaborar modelos bastante precisos con resultados que superan, incluso, trabajos de referencia de la actualidad. El modelo con mejor resultado logra valores > 0.91% en F1 Score y de >96% en exactitud de media en todos los subtipos. Este modelo es capaz de clasificar correctamente con unos niveles altos de precision pudiendo llegar a ser un sistema de soporte para médicos y radiólogos en hospitales.

Palabras Clave: Aprendizaje Profundo, Análisis de Imagen Médica, Hemorragia Intracraneal, Preprocesamiento de Imagen, Redes Neuronales Convolucionales.

Abstract

In the last two decades, the medical world has been taking a remarkable interest in the use of image processing with Artificial Intelligence for the detection of numerous diseases and injuries that occur in humans. Specifically, the techniques of Machine Learning and Deep Learning have shown their benefits to reduce human error and support more accurate diagnoses than those obtained through traditional methods by trying to find hidden patterns within the images to be processed. In this Master's Thesis, different image pre-processing techniques and training algorithms are applied for the detection and identification of different types of

Intracreneal Hemorrhage (ICH) through computed tomography (CT). All this is implemented in an experimental setting with the aim of knowing which technique and algorithm offers the best result. The results show that the use of convolutional neural networks, pre-processing techniques such as Windowing and the use of data balancing methods allow the elaboration of quite precise models with results that surpass even current reference works. The best performing model achieves values > 0.91% in F1 Score and >96% in average accuracy for all subtypes. This model can classify ICH and its subtype with a high-level accuracy and can become a support system for physicians and radiologists in hospitals.

Keywords: Deep Learning, Medical Image Analysis, Intracranial Hemorrhage, Image Preprocessing, Convolutional Neuronal Networks

Capítulo 1. Introducción

Esta sección inicia con la motivación de llevar a cabo esta investigación, partiendo del impacto que tiene la detección rápida de las Hemorragias Intracraneales y el impacto que tiene el uso de la Inteligencia Artificial aplicada al mundo de la medicina. Seguidamente, se presenta la problemática específica a resolver con esta investigación. Finalmente, se realiza una capitulación de cada uno de los apartados que contiene este documento.

1.1 Motivación

Las Hemorragias Intracraneales, en inglés Intracranial Hemorrhage (ICH), son un problema de salud grave originados por una lesión vascular provocada en el interior del cerebro. La lesión puede venir originada porque un vaso sanguíneo se rompe dejando salir gran cantidad de sangre y dañando al cerebro (Lacerda Gallardo, 2000).

Este tipo de dolencia afecta cada año a más de 130.000 españoles, de las cuales la mitad sufren alguna limitación en sus funciones a causa de las ICH. Es la segunda causa de muerte en nuestro país, la primera entre las mujeres (Díaz-Guzmán J, 2008).

Cuando un sujeto presenta síntomas de un accidente cerebrovascular, es de vital importancia la rápida intervención del personal médico para realizar un examen médico e intervenir lo más rápidamente posible. De lo contrario, graves secuelas podrían derivarse al enfermo e incluso el resultado podría ser fatal llegándose a producir la muerte.

Por ello, con el fin de evitar dar lugar a estas situaciones, se requiere un reconocimiento inmediato para determinar un intensivo y adecuado tratamiento médico para los pacientes que acuden a los centros hospitalarios con síntomas agudos de accidente cerebrovascular. Especialmente durante la noche, los fines de semana o en los países en desarrollo, en muchos de los centros de salud, urgencias u hospitalarios no tienen acceso a personal altamente cualificado como los neurorradiólogos. Desgraciadamente, y debido a esto, en muchas ocasiones es el personal no experto en la materia quien diagnostica u omite la hemorragia aguda.

En contexto, la Inteligencia Artificial, específicamente con el desarrollo del Deep Learning, han permitido el desarrollo de sistemas para la detección de ICH a partir de imágenes médicas. Los estudios indican que se han logrado resultados prometedores (Ye, 2019) (Arbabshirani, 2018) en cuanto a que pueden apoyar el diagnóstico de los especialistas, aportando más evidencias. Todo ello se logra, considerando que se pueden identificar ciertos patrones de manera rápida y precisa, tal como se requiere en el caso de las ICH. La importancia de la disponibilidad en los centros de salud de una segunda opinión confiable, "virtual" y de bajo coste, que es entrenada por neurorradiólogos y que puede ayudar a que los servicios médicos expertos y no expertos sean más eficientes y confiados.

Además, es muy importante conseguir un sistema altamente eficiente y sobre todo fiable ya que contribuiría a descargar notablemente la carga de trabajo de los hospitales. Un sistema artificial que podría actuar como un primer filtro de las tomografías computarizadas a analizar y en caso de que éstas presenten hemorragias, alertar al equipo de atención altamente cualificado.

Por lo tanto, entendiendo la importancia de un rápido diagnóstico de las ICH junto a las razones que acabamos de explicar, este trabajo se centró en la detección de las ICH con el fin de agilizar su diagnóstico y mejorar los sistemas y procedimientos hoy día existentes en el sector médico.

1.2 Planteamiento del Trabajo

Especialmente, uno de los campos donde la Inteligencia Artificial más está trabajando es el de la medicina, principalmente, en la automatización de procesos realizados tradicionalmente por el personal clínico. Una de las tareas dónde el Aprendizaje Automático destaca es, por ejemplo, la automatización del proceso documental relacionados con los diferentes exámenes clínicos a través de sistemas cognitivos capaces de manejar enormes cantidades de información para la ayuda en la toma de decisiones de los profesionales médicos. Otra de las tareas recurridas en el uso de Machine Learning en el mundo de la medicina es la identificación de patrones ocultos dónde el ser humano es incapaz de detectar por sí mismo de una manera rápida y precisa. Por ejemplo, como es el caso de esta investigación, la detección de hemorragias intracraneales a través de imágenes computarizadas del cráneo con el fin de obtener un diagnóstico médico.

Por lo tanto, se puede ver como la IA aplicada a la medicina es capaz de ayudar a mejorar la calidad de vida de las personas e, incluso, a salvar sus vidas detectando, prediciendo o, incluso, anteponiéndose a muchos de los problemas que podrían afectar a un paciente en un futuro próximo.

A pesar de que aún queda un largo camino, los grandes avances ocurridos en los últimos años dentro de la disciplina del Deep Learning se brinda la posibilidad de ser capaces hoy día de obtener mejoras significativas en la identificación, diagnóstico y tratamiento de numerosas patologías, así como infinidad de otras aplicaciones en este ámbito.

Por ello, este Trabajo de Fin de Master se ha centrado en el uso de Deep Learning para la detección automatizada de ICH siendo capaz de la identificación del tipo de hemorragia que pudiera estar presente. La mayoría de los modelos que se hacen mención en esta investigación consiguen tener un gran potencial usando diferentes tipos de técnicas y métodos de inteligencia artificial que, incluso, podrían llegar a ser desplegados en entornos clínicos reales. Por tanto, es importante preguntarse, ¿cuál de las opciones ofrece mejor rendimiento, no sólo a la hora de desplegar sistemas en entornos reales sino también a la hora de profundizar en la investigación? ¿qué técnicas, de las disponibles, logra los mejores resultados? Estas son unas de las preguntas de investigación que han orientado este TFM y que se intenta resolver tratando de encontrar cuales son las mejores técnicas para conseguir mejorar la capacidad predictiva del modelo.

1.3 Estructura del Trabajo

A continuación, se describen a grandes rasgos cada uno de los capítulos que componen este estudio:

En el presente capítulo, capítulo 1, se expone la introducción de la investigación, tratando los siguientes temas:

- Motivación, donde se trata de justificar la importancia de nuestro estudio.
- Identificación del problema a tratar, así como las causas que lo provocan.
- Planteamiento del trabajo, donde se presentan los objetivos generales (a grandes rasgos).
- Adelanto de la contribución que se pretende con este trabajo.

- Y, finalmente, un resumen de forma esquemática explicando lo esencial de cada capítulo para presentar como está estructurado este documento.

Posteriormente, en el capítulo 2, se detalla el contexto y el estado del arte dónde se tratan los siguientes temas:

- Conceptos técnicos: aquí se profundiza en los conceptos sobre los que se va a sostener este TFM. Principalmente, se abordarán temas relacionados con:
    - Deep Learning y Redes Neuronales, donde se hace hincapié en los conceptos sobre los que se sostienen estas tecnologías y las diferentes técnicas y redes neuronales profundas más interesantes.
    - Preprocesado de Imágenes, donde se describen las diferentes técnicas usadas hoy día (Windowing, Resizing, Normalization, …)
    - Métodos de Balanceo de clases y Técnicas de regularización, donde se describen las técnicas que ayudan a resolver el sobre ajuste del modelo (Undersampling, Oversampling, Data Augmentation, …)
- Conceptos clínicos: en este punto se describe a fondo el marco teórico sobre el diagnóstico de las ICH a nivel médico.
- Revisión del estado del arte: en este apartado se lleva a cabo un proceso de revisión de la bibliografía del diagnóstico de las ICH usando técnicas de Inteligencia Artificial.
- Valoración final sobre el contexto y estado del arte, lugar donde se expone por dónde se decide atacar a este problema en base a la literatura expuesta en este capítulo.

Más adelante, en el capítulo 3, se detallan los objetivos generales y específicos que se han perseguido en este TFM y la metodología de trabajo que se ha utilizado para conseguir lograr los objetivos.

A partir de aquí, se contempla toda la fase comparativa experimental de los diferentes modelos a analizar.

En el capítulo 4, se presenta la fase comparativa de esta tesis, se inicia con el planteamiento de la comparativa que se ha llevado a cabo en esta investigación, indicando qué soluciones se van a evaluar y porqué.

Seguidamente, en el capítulo 5, se continua con el desarrollo de la comparativa, donde se exponen los detalles, se describen los diferentes modelos implementados y se resumen brevemente los resultados obtenidos.

Posteriormente, en el capítulo 6, se realiza una discusión y un análisis pormenorizado de los resultados obtenidos en cada una de las pruebas, contextualizando con los obtenidos en otras investigaciones similares. Además, se emiten las recomendaciones oportunas para la implementación de modelos de este tipo, a partir de las conclusiones del análisis anterior

Y finalmente, en el capítulo 7, se clausura este estudio con las conclusiones extraídas del trabajo. Además, se incorporan las líneas de trabajo futuras que han quedado pendientes a partir de los resultados de esta tesis.

Capítulo 2. Contexto y Estado del Arte

A continuación, se presenta una descripción del papel que juega el Aprendizaje Profundo destacando los conceptos teóricos de las principales técnicas de entrenamiento principales investigaciones. Posteriormente, se expone la investigación realizada directamente sobre el campo de la medicina con el fin de obtener los conceptos clínicos de las ICH. Seguidamente, se continúa exponiendo el estado del arte de los principales trabajos que se han realizado hasta la fecha en relación con la detección de ICH en el mundo de la IA donde se explican las diferentes técnicas usadas en los mismos. Por último, se extraen las conclusiones de la revisión de la bibliografía que se ha expuesto en este capítulo.

Todo el proceso de revisión de bibliografía se ha llevado a cabo buscando las publicaciones más relevantes del estado del arte. Principalmente, la herramienta base utilizada ha sido el motor de búsqueda de Google Scholar, aunque también se ha utilizado otros buscadores de artículos académicos y científicos como ResearchGate.

2.1 Conceptos Técnicos

En este apartado se presenta el marco teórico a nivel técnico relacionado con las redes neuronales profundas. Una red neuronal es un problema complejo de optimización con muchos parámetros en el que obtener una solución adecuada no es trivial. Por ello, a continuación, aparte de explicar qué son las redes neuronales, se presentan las diferentes técnicas y métodos que existen hoy día para poder entrenarlas de una manera más efectiva y eficiente a fin de obtener los mejores resultados para resolver problemas de clasificación con imágenes, tal y como puede ser el diagnóstico de ICH.

2.1.1 Deep Learning y Redes Neuronales Profundas

El Aprendizaje Profundo, o Deep Learning en inglés, es un área dentro de la Inteligencia Artificial que busca construir automáticamente conceptos complejos a partir de conceptos sencillos. Dentro de las diferentes técnicas que ofrece el Deep Learning la más relevante es red neuronal (Schmidhuber, 2015). Básicamente, una red neuronal es una función matemática definida por nodos. Estos nodos están distribuidos en capas donde cada nodo es, a su vez, una función aún más simple que depende de los nodos de la capa anterior. De esta manera, cada nodo utiliza los conceptos simples aprendidos en capas anteriores para desarrollar un concepto más complejo a partir de ellos (Nielsen M. A., 2015). Dentro de cada nodo las

entradas tienen asignados unos determinados pesos "wi" y unos biases b (Goodfellow, 2016). En la Ilustración 1 se visualiza como se obtiene la expresión matemática.

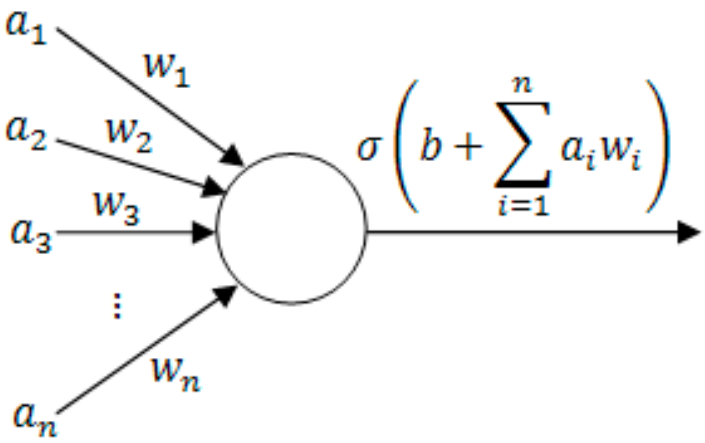

Ilustración 1 Expresión matemática de una neurona en una red neuronal (Goodfellow, 2016).

La función de activación (σ) se conoce también como *non-linearity,* es una función diferenciable que permite introducir cierta no-linearidad en las secuencias tratadas por la red neuronal. La capa de entrada recibe el input (texto, imágenes, audio, etc.), una o más capas ocultas procesan la información, y finalmente una capa de salida emite el resultado (Nielsen M. , 2019). Estas redes se entrenan de forma supervisada, por medio un algoritmo conocido como backpropagation. Este consiste en ir modificando los pesos w y los biases b en sucesivas iteraciones (batches y epochs), buscando minimizar una función de coste. Se apoya en una técnica conocida como descenso del gradiente (gradient descent), y el concepto de las derivadas parciales (LeCun, 1998).

Las redes neuronales son modelos con un gran poder de representación. En la Ilustración 2 se muestra un pequeño ejemplo de cómo es posible que una red neuronal es capaz de aprender conceptos abstractos o de alto nivel a partir de conceptos simples tales como contornos o bordes. Esto permite que la red neuronal sea capaz de aprender una representación de los datos adecuada consiguiendo finalmente clasificar la fotografía como un animal, una persona o un objeto como un coche.

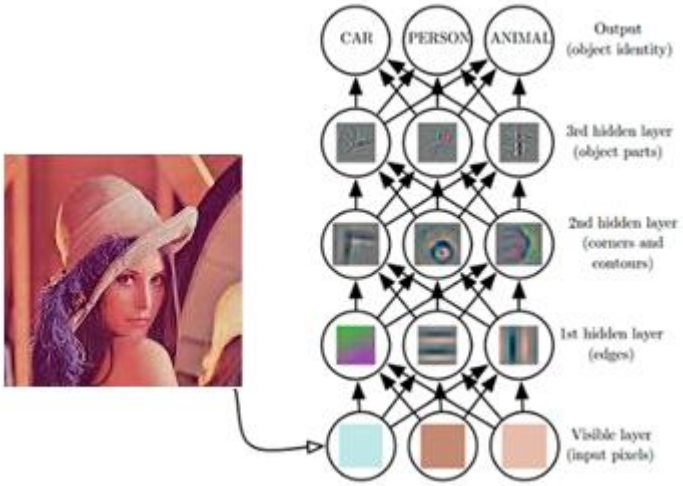

Ilustración 2 Ejemplo de una red neuronal aprendiendo conceptos complejos (Goodfellow, 2016).

Ante un problema de clasificación de imágenes como el que se muestra en la citada Ilustración, una de las maneras de resolverse puede ser analizando pixel por pixel, sin embargo, este es un trabajo muy costoso computacionalmente y, más aún, cuando es necesario procesar imágenes de un tamaño considerable. Para evitar esto, se utilizan una de las técnicas más prometedoras y que mayores avances han aportado en los últimos años: las redes neuronales convolucionales, o convolutional neural networks (CNN) en inglés. Este tipo de redes son la mejor herramienta que existe hoy día ya que son las que mejor resultado proporcionan hoy día para los problemas de procesamiento de imágenes (Krizhevsky, 2012).

Las redes convolucionales son muy similares a las redes neuronales convencionales con la peculiaridad de que incorporan una serie de características y elementos que se utilizan para el procesado de la imagen. En la Ilustración 3 se puede ver un ejemplo gráfico de cómo es la arquitectura de una red neuronal con capas convolucionales.

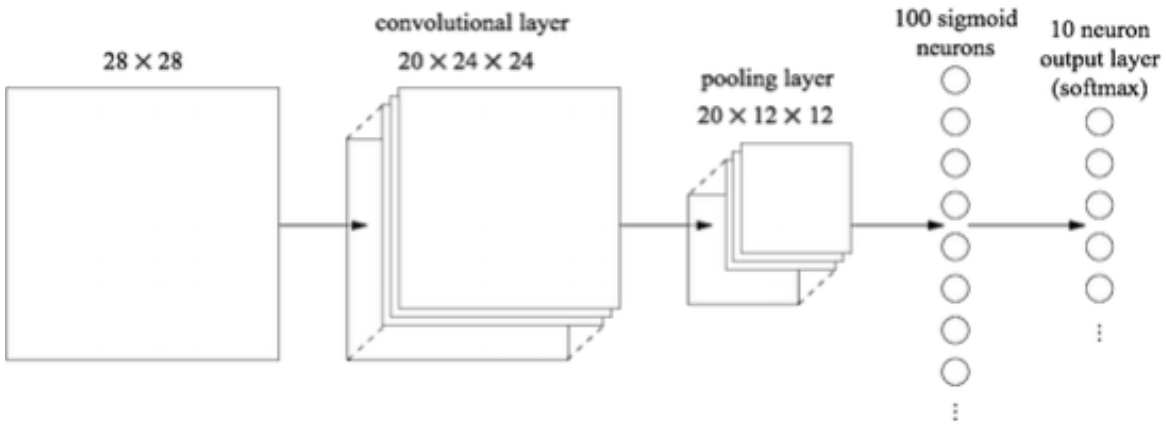

Ilustración 3 Arquitectura de una CNN (Nielsen M. A., 2015)

Existen todos tipos de capas: las capas convolucionales y las capas pooling (Rawat, 2017). En cuanto a las capas convolucionales analizan la imagen de entrada haciendo pasar un filtro convolucional que va recorriendo la imagen, obteniendo un valor por cada pixel sobre el que se aplica la convolución. Estas capas también contienen parámetros que la red debe de ajustar en base a lo que va aprendiendo, tal y como en las redes neuronales tradicionales. Por otro lado, en cuanto a las capas pooling, son las encargadas de reducir las dimensiones de las imágenes tratadas generando una salida más reducida espacialmente, pero a la vez más profunda. También consiste en aplicar una ventana deslizable, como el filtro convolucional, pero escogiendo el valor más alto de la ventana (capas max pooling) o la media (capas average pooling), dependiendo de la estrategia a seguir. Básicamente, la red neuronal convolucional es una consecución de capas convolucionales y pooling que a medida que va avanzando la imagen por ella se va generando una representación cada vez más pequeña espacialmente pero más profunda donde se obtiene una jerarquía de conceptos abstractos sobre la imagen. Esta jerarquía normalmente es utilizada como un extractor de características reaprovechado como entrada a una fully connected layer (o capa completamente interconectada), como se aprecia en la Ilustración 3, donde el número de neuronas utilizadas en esas capas fully connected es infinitamente menor que en el caso de haber procesado la imagen pixel a pixel.

A lo largo de esta última década, se ha avanzado bastante en el mundo de las redes convolucionales llegando a dejar prácticamente resuelto muchos de los problemas de clasificación de imágenes. A continuación, se resumen las redes convolucionales que más impacto han producido en los últimos años. Todas estas redes han sido entrenadas con ImageNet y cada una, en su momento, fue la ganadora de la competición ILSVRC («Large Scale Visual Recognition Challenge», 2015):

- AlexNet (Krizhevsky, 2012). Esta red marcó un punto de inflexión en la historia de la visión por computador. Fue pionera en el uso de técnicas como las unidades ReLU en vez del uso de las tradicionales unidades de activación como sigmoide. Además, también fue la primera en el uso de GPUs para el entrenamiento de este tipo de redes.

- VGGNet (Zisserman, 2014). A partir de esta red se aumentó notablemente la profundidad. En este caso, esta red tenía 19 capas. Otra de las caracterizas es que utilizaba filtros convolucionales más pequeños, de 3x3 píxeles. Aparecieron conceptos nuevos como el Transfer Learning, que permitieron realizar fine-tuning de una red previamente entrenada en ImageNet.

- Inception (Szegedy, 2015). Esta red se caracterizaba por aumentar su profundidad y anchura, llegando a las 22 capas. Buscaba mejorar la eficiencia del entrenamiento, con el fin de que su procesamiento fuera más ligero para entornos de tiempo real, smartphones o sistemas embebidos.

- ResNet (K. He, 2015). La novedad de esta red de la mano de Microsoft fueron los bloques residuales, que simplificaban el entrenamiento de redes muy profundas. Crearon una red residual de 152 capas, 8 veces más profunda que VGGNet, y pese a ello menos costosa de entrenar.

- InceptionV3 (V. Vanhoucke, 2016). Es una versión mejorada la Inception que se acaba de presentar con el mismo objetivo, la eficiencia.

De todas estas redes, las más utilizadas y aquellas que mejor resultado están proporcionando son las basadas en arquitecturas ResNet e InceptionV3. Incluso, la mayoría de frameworks, como Keras (Chollet, 2015) que incorporan librerías para el desarrollo de redes neuronales, proporcionan estos modelos previamente entrenados con ImageNet para servirlos o para hacer reajustes o fine-tuning con ellas. Esta técnica, que también puede ser reaprovechada para el procesamiento de imágenes médicas, consiste en aprovechar el ajuste previo de los parámetros fruto del entrenamiento previo de la red neuronal y terminar de reajustarlos en un área de aplicación en concreto. De esta manera, se reaprovecha todo el conocimiento que tiene una red y se adapta para poder resolver una tarea para la cual no fue entrenada la red en un principio. El uso de esta técnica a día de hoy está muy extendido: para imágenes biomédicas (Zhou, 2017), para reconocimiento de patrones de escritura a mano (Rosa, 2015), para reconocimiento de plantas (Reyes, 2015), …. Idealmente, para obtener los mejores resultados, la red debe estar previamente entrenada con imágenes del mismo tipo de las que se va a utilizar para hacer el reajuste, es decir, para el diagnóstico de ICH, se puede utilizar una red previamente entrenada con un dataset de tomografías computarizadas del cerebro para estimar, por ejemplo, la cantidad de materia gris y la anatomía del mismo. Sin embargo, debido a la complejidad de estas redes, muchas veces no es posible tener una cantidad de datos ingente para conseguir un entrenamiento óptimo de la misma, sobre todo, cuando se habla de imágenes médicas.

Por esto, aunque las imágenes del conjunto de datos de ImageNet es un conjunto muy dispar respecto a las tomografías computarizadas del cerebro que se usan en este trabajo, esto es

un inconveniente mucho mayor a ojos de un humano que para una CNN, que es capaz de extraer patrones ocultos que un humano no es capaz de identificar de manera sencilla.

Dentro de la aplicación de reajuste de redes convolucionales sobre imágenes médicas, se pueden identificar estudios donde el uso de estas técnicas ha servido para clasificar exitosamente más de 2000 variantes de cáncer de piel con una red InceptionV3 (Esteva, 2017), para lograr detectar retinopatía diabética con una red InceptionV3 (Gulshan, 2016) y para realizar el diagnóstico de cáncer de mama con muy buenos resultados (Vesal et al., 2017) comparando una con una red InceptionV3 con una ResNet50. Dentro de mundo del diagnóstico automático de ICH se puede identificar uno de los trabajos más recientes y relevantes donde se utilizan diferentes tipos de CNN para clasificar incluso el subtipo de ICH (Ye, 2019) o el trabajo de Arbabshirani que consistía en la utilización de CNN para la identificación de ICH y la optimización de las listas de trabajos (Arbabshirani, 2018).

2.1.2 Procesado de las imágenes médicas

Técnicas como la Tomografía Computarizada (TC) y la Resonancia Magnética (RM) han revolucionado el mundo de la medicina desde su descubrimiento alrededor de los años 60 y 70. A través de la obtención de imágenes médicas, éstas han permitido el estudio, análisis y visualización de las estructuras internas de todo el cuerpo humano. Estas técnicas facilitan a los especialistas el diagnóstico de enfermedades y todo tipo de anomalías en el interior del organismo (Kidwell, 2004).

Para poder construir un modelo de Aprendizaje Automático correctamente es muy importante realizar un buen procesamiento previo de los datos. Existe una gran variedad de técnicas de procesamiento de imágenes, cada una con diferentes objetivos, tales como: suavizado, realce, extracción de regiones o bordes, descripción de líneas y contornos o regiones, reconocimientos de formas, … todas con el fin común de ayudar a adaptar las imágenes para ser aprovechadas de una manera más eficiente y efectiva por la red o el modelo en cuestión. Debido a la extensión del campo, en este trabajo, se explican tres de las técnicas de procesado más utilizadas con imágenes médicas: Windowing, Resizing y Normalización, y que serán objeto de estudio en este proyecto.

2.1.2.1 Windowing

Para poder explicar esta técnica es necesario entender que hay detrás de las Tomografías Computarizadas (TC), o en inglés, Computed Tomography. La TC es una tecnología de exploración de rayos X que permite obtener imágenes detalladas de alta resolución de diferentes cortes a lo largo de todo el cuerpo. A diferencia de la radiografía convencional, la TC obtiene una gran variedad de imágenes al incidir rayos X alrededor del cuerpo y generar imágenes desde diferentes ángulos de este. Posteriormente, una computadora combina todas estas imágenes en una imagen final que representa un corte del cuerpo como si fuera una rodaja, también llamado corte (o slice en inglés) reconstruyendo una imagen plana del segmento. Finalmente, se obtienen múltiples imágenes de diferentes cortes del cuerpo, logrando tener múltiples vistas de la anatomía a estudiar (Buzug, 2011).

Normalmente, la reconstrucción final suele obtenerse en cualquier de los tres plano o cortes anatómicos: axial, coronal y sagital. Se muestra un ejemplo en la Ilustración 4.

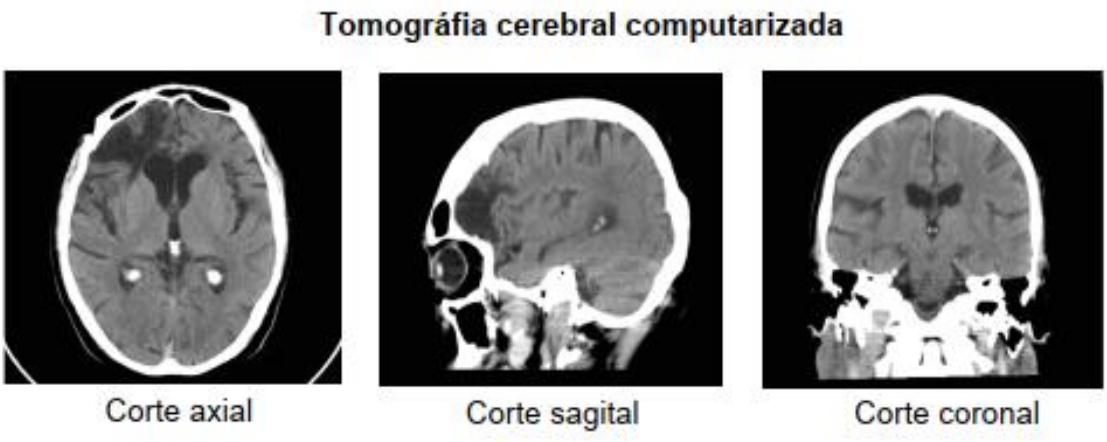

Ilustración 4 Cortes axial, sagital y coronal en una tomografía cerebral (Martínez, 2011)

La TC se basa en los principios de la ecuación que indican la atenuación que sufre un haz de rayos X al atravesar un elemento en concreto:

$$I_x = I_0 \cdot e^{-\mu x}$$

Donde I_x es el valor de la intensidad de los rayos X después de atravesar el espesor *x* de un objeto. Por otro lado, I_0 es el valor de la intensidad de la radiación que incide sobre la superficie de este. Por último, *e* es el número base del sistema logarítmico natural y el coeficiente lineal de atenuación μ depende del número atómico, de la densidad del medio y de la longitud de onda del haz de radiación incidente (C.R. Becker, 2008).

De esta manera, es posible conocer la densidad de los tejidos debido a que esta densidad es directamente proporcional a la atenuación de energía absorbida por el tejido tras proyectar rayos X sobre el cuerpo. Una vez reconstruida la imagen de TC se obtiene una matriz con los niveles de absorción a cada volumen estudiado, llamado vóxel, el cual se representa en un solo plano como un píxel. Si se trata de visualizar una TC, cada píxel se puede apreciar con un nivel concreto de brillo, sin embargo, este se mantiene almacenado de manera digital como un coeficiente de atenuación lineal μ (Fraile, 2004).

A través de estos coeficientes de atenuación se obtiene la capacidad de saber qué material ha sido atravesado por el haz de rayos X. Con el fin de evitar discrepancias en la obtención de estos coeficientes originadas por los diferentes equipos de medición que hay en el mercado, se ha adoptado universalmente el uso de una escala común llamada escala de Hounsfield. Básicamente, esta escala establece que se realizar una transformación lineal que permite en contra números de TC dentro de una escala común, escala que establece un valor de TC de 0 para el agua y -1000 para el aire (Martínez, 2011).

$$Número\ de\ TC = \frac{(\mu_{\text{material}} - \mu_{\text{agua}}) \cdot E}{K}$$

En esta ecuación, E es la energía efectiva del haz de rayos X. Por otro lado, μmaterial representa el coeficiente lineal de atenuación del material que se atraviesa. Y finalmente, μagua contiene el coeficiente de atenuación del agua y K es una constante que depende del diseño del equipo (Fraile, 2004).

Esta escala establece una serie de valores para cada pixel, también llamados unidades de Hounsfield, que oscilan desde los -1000HU y los +1000HU. Por normal general, sólo se representa mediante una escala de grises un sector parcial de los valores de la tomografía con el fin de visualizar únicamente los detalles del órgano o tejido a estudiar. Esta técnica es la llamada Ventana de TC o Windowing y es debido a que, en el caso de asignar a cada

unidad un nivel de brillo diferente, un radiólogo no sería capaz de distinguirlos ya que el ojo humano no distingue más de 40 tonalidades de brillo diferentes. Por ello, representar en una imagen toda la gama de valores de la escala de Hounsfield supondría no ser capaces de poder visualizar una gran cantidad de información importante para el diagnóstico (Martínez, 2011).

La utilización de ventanas permite extraer la información relevante de la propia tomografía de una parte especifica que es la que se quiere visualizar (Prieto, 2005).

Algunas de las más utilizadas por parte de los radiólogos se muestran a continuación:

- Brain window o ventana de cerebro: centrada en 40HU y ancho de 80HU
- Blood/Subdural window o ventana de subdural o sangre: centrada en 50-100HU y ancho de 130-300HU
- Soft tissue window o ventana de los tejidos blandos: centrada en 20–60HU y ancho de 350–400 HU
- Bone window: centrada en 600HU y ancho de 2800HU

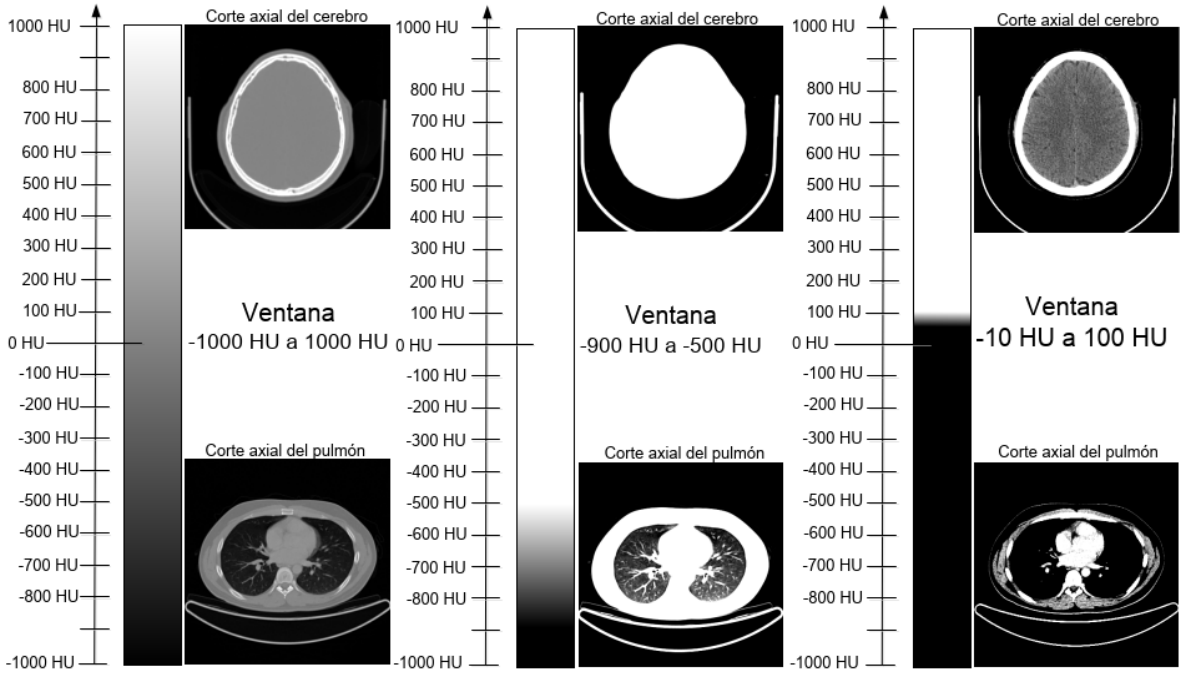

Ilustración 5 Visualización en un monitor de video una tomografía computarizada con diferentes ventanas. La selección de una ventana u otra permite poder visualizar los detalles dependiendo del tipo de tejido. (Martínez, 2011)

En la Ilustración 5 se puede apreciar un ejemplo gráfico de la aplicación de diferentes tipos de ventanas y su resultado.

Otros tipos de ventanas también utilizados en entornos médicos pueden ser:

- Ventana de grasas: centrada en -100HU y ancho de -50HU
- Ventana de tejido muscular: centrada en 40HU y ancho de 60HU
- Ventana de tejido arterial: centrada en 10HU y ancho de 100HU
- Ventana de medula espinar: centrada en 150HU y ancho de 200HU

2.1.2.2 Resizing

Esta técnica es una de las técnicas más comunes cuyo objetivo consiste en adaptar el tamaño de las imágenes de dataset en base a una referencia con el fin de que puedan ser aprovechados de la mejor forma posible por los modelos buscando que los píxeles de todas representen las mismas estructuras anatómicas (Woods, 1993) (Klein, 2009).

Cuando las imágenes del dataset tienen tamaños diferentes unas entre otras, esta técnica tiene mucho sentido ya que establece un tamaño común para que los modelos puedan centrarse en determinadas regiones de las imágenes, teniendo la seguridad de que en todas ellas se representa un mismo elemento anatómico.

Por otro lado, es muy común redimensionar las imágenes de entrada al modelo con el fin de entrenar más rápidamente el modelo. Por ejemplo, en el trabajo de Krizhevsky (Krizhevsky, 2012) en un problema de clasificación donde se construyó AlexNet, se partió de un conjunto de datos con imágenes de alta resolución y se ajustaron todas las imágenes a una dimensión 256x256. Se podría haber mantenido el valor inicial de las imágenes, sin embargo, la red tardaría mucho más en entrenar. También, se podría establecer a valores más pequeños, pero esto podría generar problemas ya que si los detalles a detectar en las imágenes son muy pequeños podría ser ocultados y afectar en la precisión del modelo.

2.1.2.3 Whitening o Normalización

Otra de las técnicas que más se utilizan a la hora de procesar imágenes en Aprendizaje Automático es Whitening o normalización de la intensidad de cada imagen. El objetivo de esta técnica es hacer que los valores de los pixeles de las imágenes del conjunto de datos con el que se va a alimentar nuestra red estén todos en una escala común, sin afectar o distorsionar las diferencias entre los rangos de valores originales (Sudeep, 2017). Un ejemplo podría ser

reducir el valor de la intensidad de cada pixel en valores reales entre 0 y 1 resultante de restar la media y dividirlo por la desviación típica.

$$\frac{data - (data)}{\max(data) - \min(data)}$$

Por último, es importante recalcar que no todos los conjuntos de datos requieren normalización, únicamente se requiere en aquellos donde las características de los datos (en el caso de las imágenes médicas, los pixeles) tengan rangos dispares (L. Wan, 2013).

2.1.3 Técnicas de Balanceo de Clases y Regularización

Cuando se trata de resolver un problema de clasificación a partir de un conjunto de datos etiquetado con imágenes médicas es muy probable que las clases del conjunto estén desbalanceadas. Esto quiere decir, que existe un desequilibrio entre las imágenes etiquetadas como positivas (por ejemplo, que presentan la enfermedad a detectar) y las negativas. Este es un problema muy común que ocurre en Aprendizaje Automático que sobre todo sucede en problemas de detección de fraude, diagnóstico de enfermedades, filtro de spam, …

En cuanto a los datos médicos, que es el objeto de esta tesis, suele ser mucho más sencillo encontrar información de pacientes no enfermos que de pacientes enfermos lo que implica que la clase negativa suele estar fuertemente correlacionada, mientras que existe muchísima variación en la clase positiva (Greenspan, 2016)

A continuación, se exponen las técnicas que se utilizan hoy día para solventar este problema tan común en el mundo del Aprendizaje Automático.

2.1.3.1 Undersampling

Esta técnica implica eliminar de manera aleatoria las observaciones de la clase mayoritaria para evitar que su señal domine el algoritmo de aprendizaje. La heurística más común para hacerlo es el remuestreo sin reemplazo. Se suele separar las observaciones de cada clase en diferentes conjuntos de datos, por ejemplo, observaciones positivas y observaciones negativas. A continuación, se remuestrea la clase mayoritaria sin reemplazo, estableciendo el número de muestras con el fin de que coincida con el de la clase minoritaria. Por último, se combinan ambos conjuntos de datos componiendo un conjunto de datos balanceado al 50%.

Un buen ejemplo puede ser el trabajo de Mazurowskia (Maciej A. Mazurowskia, 2008) que demostraron que el desbalanceo de clases afecta muy negativamente a la precisión del modelo. Sin embargo, llegaron a la conclusión de que eliminar observaciones de la clase con más representación casi nunca es bueno.

2.1.3.2 Oversampling

Esta técnica es totalmente contraria a la anterior. Esta técnica es el proceso de duplicar aleatoriamente las observaciones de la clase minoritaria para reforzar su señal. Existen varias maneras de conseguirlo, pero la más común es simplemente remuestrear con reemplazo.
En este caso, lo que se suele hacer es separar las observaciones de cada clase en diferentes conjuntos de datos. Después, se remuestrea la clase minoritaria con reemplazo de tal manera que se establece el número de muestras para que coincida con la de la clase mayoritaria. Por último, se combinan ambos conjuntos de datos componiendo un conjunto mucho más grande que el de undersampling y balanceado también al 50%. En la Ilustración 6, se puede ver un ejemplo de como un modelo entrenado con oversampling es capaz de poder clasificar la tercera clase cuando en el dataset original no fue capaz.

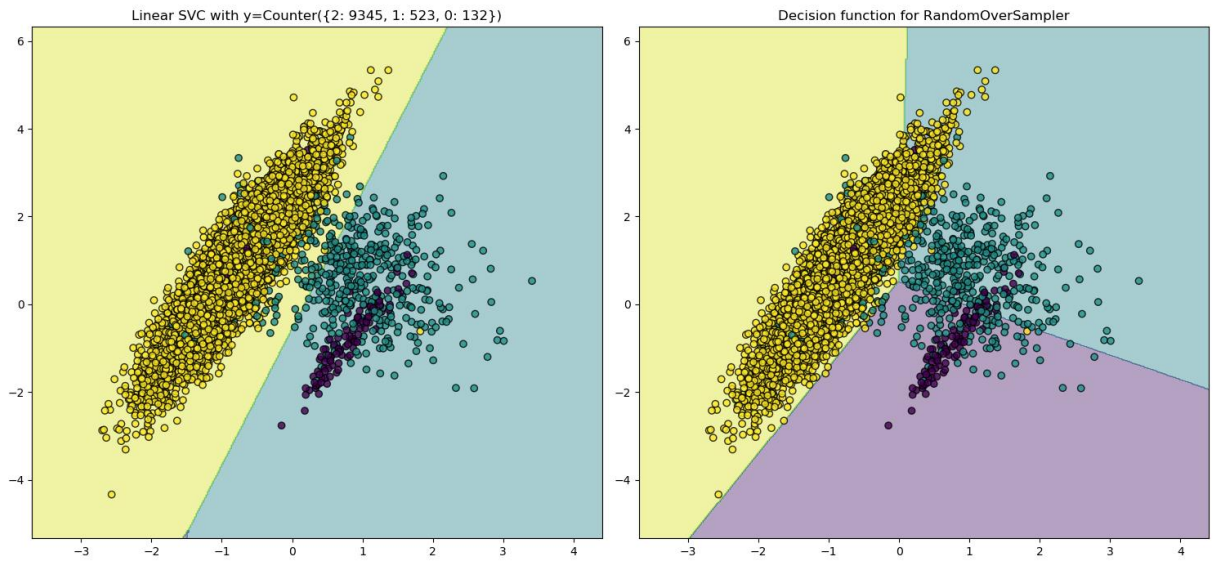

Ilustración 6 Comparación de las funciones de decisión resultantes de un clasificador entrenado utilizando el conjunto de datos original (izquierda) y el conjunto de datos aplicando oversampling (derecha). (Nogueira, 2016)

En este caso, en el estudio de (Maciej A. Mazurowskia, 2008), demostraron que en determinadas situaciones duplicar observaciones de la clase con menor representación puede mejorar los resultados muy ligeramente. Sin embargo, al ser tan pequeña la mejora, el oversampling no ha llegado a cobrar demasiada importancia en la bibliografía.

2.1.3.3 Data Augmentation

Otra de las técnicas de balanceo de datos frecuentemente usada es Data Augmentation (DA) o técnicas de aumento de datos. Se suelen utilizar cuando el conjunto de datos no es suficiente para poder entrenar una red o cuando una (o varias) de las clases tiene una baja representación. El objetivo de esta técnica es que se consiga elevar el número de observaciones que se tienen del dataset realizando diferentes transformaciones a las observaciones reales creando observaciones sintéticas.

El proceso de obtener más datos para el conjunto de entrenamiento mediante DA puede considerarse también como una forma de regularización. Esto es debido a que, al añadir cierta aleatoriedad y heterogeneidad en las imágenes, se consigue impedir que la red aprenda ciñéndose a elementos particulares de las observaciones originales. No obstante, el uso de esta técnica debe hacerse con cuidado ya que conlleva un riesgo debido a que al estar añadiendo información transformada de muestras ya existentes puede aumentar el riesgo de overfitting, en lugar de disminuirlo que es el objetivo final, tal y como pasa con oversampling (Shin, 2018). Estas transformaciones pueden consistir en realizar desplazamientos aleatorios en anchura y altura, cortes en las imágenes, añadir zoom, alteración de brillo o contraste, imágenes simétricas... todo con el fin de aumentar el conjunto de datos de entrenamiento, (Ding, 2018). Algunas de las técnicas que permite DA se pueden apreciar gráficamente en la Ilustración ejemplo de una modificación de la imagen sería utilizar la imagen simétrica.

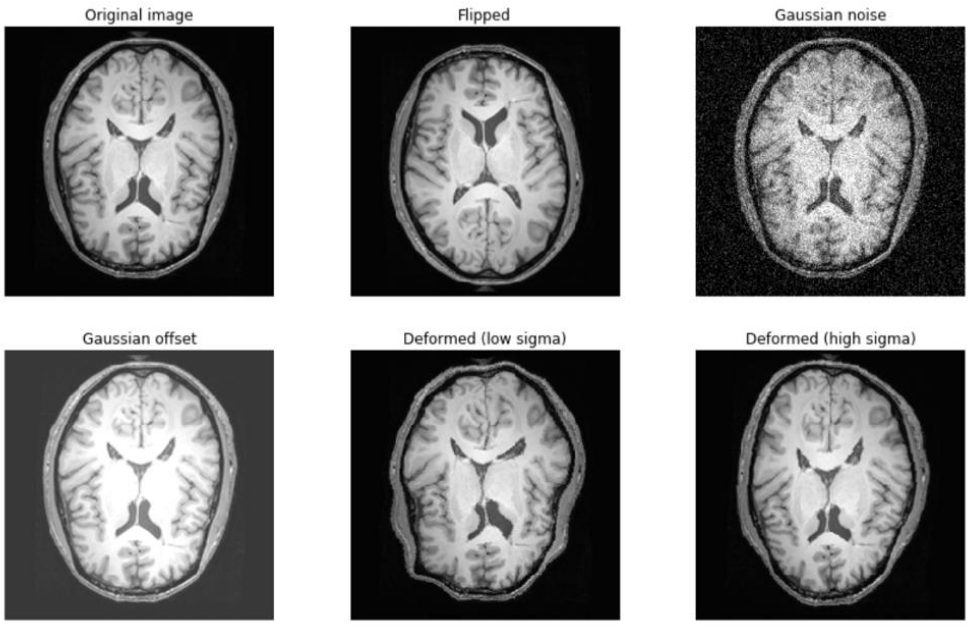

Ilustración 7 Ejemplos de técnicas de aumento espacial y de intensidad

2.1.3.4 Dropout

Las redes neuronales, como se ha explicado antes, son modelos con un gran poder de representación debido al gran numero de parámetros y capas que posee. Si la red es muy compleja y el conjunto de datos es muy pequeño, es muy posible que la red acabe aprendiendo muy bien los datos con los que está siendo entrenada llegando incluso a memorizarlos. Sin embargo, sobre un conjunto de observaciones nunca vistas es posible que no consiga establecer una generalización por lo que se obtendrá una baja capacidad de predicción en el modelo.

Este fenómeno se conoce como overfitting y es un fenómeno común en el mundo del Aprendizaje Automático. Las técnicas de regularización como dropout intentan solucionar este problema. El objetivo de esta técnica es aplicar un efecto regularizador con el fin impedir a la red memorizar resultados. Es una técnica bastante moderna (Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014) y que ha encontrado una gran acogida, muy utilizada en la práctica (Darias Plasencia, 2019), (Ye, 2019), (Arbabshirani, 2018). Dropout aplica una salida con valor 0 (las desactiva) a un porcentaje de neuronas de la red durante el entrenamiento de manera aleatoria. Este porcentaje se proporciona a través de un hiperparametro y las neuronas afectadas van cambiando por cada batch. De este modo, en cada *Batch*, un número aleatorio de neuronas se desactivará. Se puede ver un ejemplo gráfico en la Ilustración

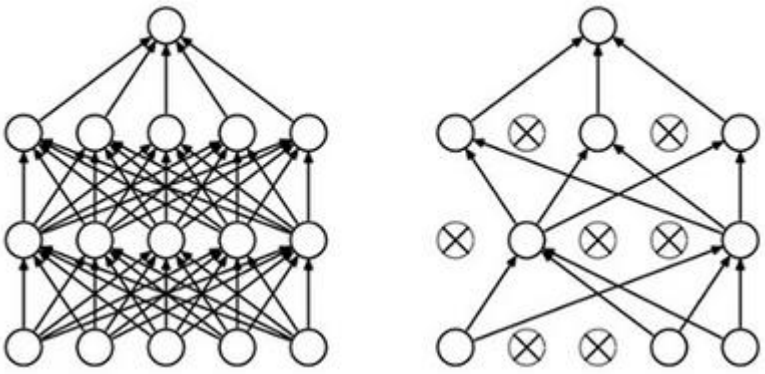

Ilustración 8 Efecto de aplicar dropout. A la izquierda se tiene una Fully Connected Network estándar y a la derecha después de aplicar un dropout del 50% (Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014)

2.2 Conceptos Clínicos

Las hemorragias intracraneales (ICH), es decir, hemorragias dentro del cráneo, son un tipo de ictus cerebral, por lo tanto, suponen un problema de salud grave que requiere la rápida intervención médica y un intensivo tratamiento médico.

Las ICH suponen el 10-15% de todos los ictus cerebrales y en nuestro país afectan cada año a 130.000 españoles. De hecho, más de 300.000 personas sufren algún tipo de limitación en sus funciones debido a esto. Es también la segunda causa de demencia después del Alzheimer y primera causa de dependencia en los adultos (Madrigal Mesa, 2016).

Cuando se produce una lesión vascular en nuestro cerebro, ésta ha podido ser producida por dos causas: cuando un vaso se obstruye, es decir, un ictus isquémico y cuando un vaso se rompe, es decir, un ictus hemorrágico. La mayor parte de las veces, se suele simplificar la forma de hablar y se refiere simplemente a ictus y hemorragia respectivamente, lo que puede llevar a confusión. Básicamente, el ictus (isquémico) es como si fuera un infarto, pero, en lugar de producirse en el corazón, se produce en el cerebro. De hecho, también se conoce como infarto o ataque cerebral. Este ataque consiste en la interrupción brusca y repentina en el flujo sanguíneo y su consecuencia significa la muerte del tejido que se encuentra alrededor del punto en el que se produce el bloqueo debido a que no pueden llegar los nutrientes necesarios ni el oxígeno. Si no se desobstruye esa zona entre 3 y 8 horas, se producirá una lesión cerebral (D. Escudero Augusto, 2008).

Por otro lado, el ictus hemorrágico, el cual es objeto de estudio en este trabajo, consiste en que una arteria dentro de nuestro cerebro se rompe dejando salir gran cantidad de sangre y dañando al mismo. La recuperación depende del tamaño de la lesión, la zona en la que se haya producido y de la velocidad de acción con que el neurólogo pueda tratar al paciente y asignarle un tratamiento efectivo (A. Qureshi, 2001).

La ICH es una condición relativamente común que tiene un sinfín de causas, entre ellas pueden ser: trauma, accidente cerebrovascular, aneurisma, malformaciones vasculares, hipertensión arterial, drogas ilícitas, trastornos de la coagulación de la sangre. Las consecuencias neurológicas también varían ampliamente dependiendo del tamaño, tipo de hemorragia y ubicación desde un simple dolor de cabeza hasta la muerte. Su localización más frecuente suele ser en la mitad de los casos en ganglios de la base, en una menor frecuencia en los lobares, y raramente en cerebelosa y tronco cerebral. En función de su localización, se

clasifican en subtipos. A continuación, se explican cada uno de los subtipos y se incorpora una ilustración de una TC de cada uno de ellos para facilitar la identificación por parte del lector:

- Epidural (EPI): se localiza entre la duramadre y el cráneo. Frecuentemente viene ocasionado por la fractura del cráneo en edades tempranas o adolescentes debido a un traumatismo cerebral.

Ilustración 9 TC de una hemorragia de tipo epidural

- Intraventricular (INV): este tipo de hemorragia se produce cuando afecta a los ventrículos del cerebro que contienen el líquido cefalorraquídeo. Afecta a niños prematuros, sobre todo a aquellos que han tenido síndrome de distrés respiratorio, colapso pulmonar o presión alta. Se localiza dentro de los ventrículos.

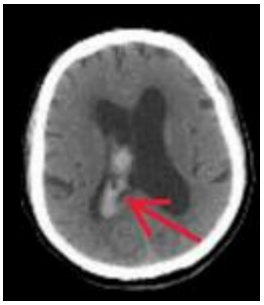

Ilustración 10 TC de una hemorragia de tipo intraventricular

- Intraparenquimatosa (INP): También conocida como Hemorragia cerebral espontánea, suele estar muy relacionada con la hipertensión arterial, problemas de coagulación, plaquetopenia, tumores, drogas, entre otros. Suele afectar a las personas mayores. Este tipo de hemorragia cerebral supone el 10-15% de todos los ictus. Se localiza dentro del cerebro.

Ilustración 11 TC de una hemorragia de tipo intraparenquimatosa

- Subaracnoidea (SAH): suele afectar a los jóvenes entre 20 y 40 años y se localiza en la zona subaracnoideo, por fuera de la piamadre. Suele estar causada por la rotura de un aneurisma, traumatismo craneal, … Es muy común en personas mayores que han sufrido alguna caída y en los jóvenes que han sufrido un accidente de tráfico.

Ilustración 12 TC de una hemorragia de tipo subaracnoidea

- Subdural (SDA): es el más común de las cinco y se localiza entre el aracnoidea y la duramadre. Las personas que lo padecen suelen caer en coma durante su hospitalización.

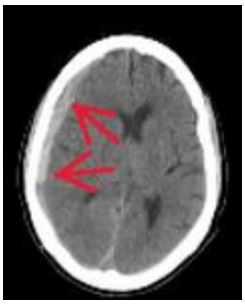

Ilustración 13 TC de una hemorragia de tipo subdural

En España, la incidencia de ICH asciende hasta los 15 casos por cada 100.000 habitantes al año, donde los varones mayores de 55 años se ven afectados en mayor frecuencia (Láinez JM, 2002). A pesar de que la ICH es bastante menos frecuente que el ictus isquémico, es la ICH la que presenta una mayor mortalidad y morbilidad, siendo una de las primeras causas

de discapacidad grave. La hemorragia cerebral no es un fenómeno monofásico que ceda inmediatamente, ya que el hematoma continúa aumentando en las primeras 24 horas (T. Brott, 1997).

A través del uso de las tomografías computarizadas (TC) se demostró que los hematomas son dinámicos en el tiempo (fig. 1). T. Brott demostró mediante TC que las hemorragias crecen, y lo hacen sobre todo en las primeras horas (26% en la primera hora y un 38% en las primeras 20 horas). Este mecanismo es el responsable del deterioro neurológico durante las primeras 24 horas. Se puede apreciar un ejemplo de su evolución a través de TC en la Ilustración 14. Por esta razón, y por las características propias de la enfermedad, es realmente crítico la detección de esta lo antes posible, de aquí el trabajo que se ha desarrollado.

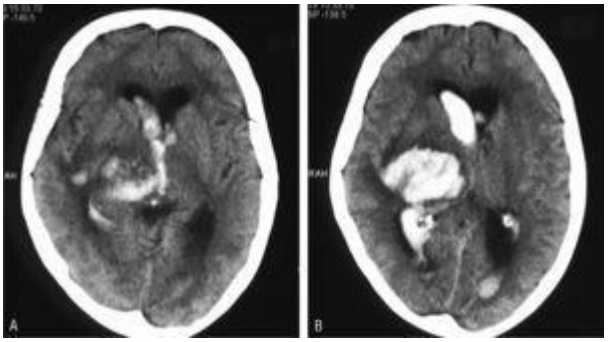

Ilustración 14 Tomografía computarizada (TC) craneal al ingreso con hemorragia derecha en ganglios de la base (izquierda). TC del mismo paciente 6 horas después que demuestra un aumento del sangrado (derecha).

Aproximadamente, el 25% de los pacientes que sufren ICH acaban teniendo un deterioro neurológico en las primeras 24 horas originado por el crecimiento del hematoma, al aumento de sangre ventricular y al edema precoz. Algo más improbable es el deterioro tardío, a veces ocurre que se produce una progresión del edema o resangrado entre la segunda y tercera semana de evolución (D. Escudero Augusto, 2008).

Entre el 35% y el 52% se estima al mes que oscila la mortalidad de la ICH, de los cuales la mitad de los fallecimientos son de manera precoz, en las primeras 48 horas por un síndrome de hipertensión endocraneal (M. Ariesen, 2003).

Por otro lado, si se trata de valorar la mortalidad al año, ésta varía según su localización. Aquellas que suceden en los ganglios basales alcanzan el 51%. Las que se producen en las lobares ronda una mortalidad del 57%. En cuanto a las que se originan en las zonas

cerebelosas el 42% y, por último, las ocurridas en la zona del tronco-encéfalo hasta un 65% (M. Ariesen, 2003).

La ICH se asocia además con una alta morbilidad. Sólo el 10% de los pacientes son independientes al mes y el 20% a los 6 meses. Hasta un 50% de los sujetos sufren algún tipo de discapacidad, lo que supone un elevado coste sanitario (D. Escudero Augusto, 2008).

El diagnóstico requiere una intervención urgente. Por ello, cuando un paciente muestra síntomas neurológicos agudos como pérdida del conocimiento o un dolor de cabeza severo, los especialistas deben inspeccionar el interior del cráneo con el fin de buscar la presencia, la ubicación y el tipo de hemorragia. Como se ha comentado en el primer capítulo, este es un proceso complicado, costoso y conlleva mucho tiempo y esfuerzo.

2.3 Revisión del Estado del Arte

El objetivo del radiólogo es detectar la hemorragia, caracterizar el subtipo, su tamaño y determinar si la hemorragia pone en peligro las zonas críticas del cerebro que pudieran requerir una cirugía inmediata. Para ello, las técnicas más usadas hoy día para el diagnóstico de las hemorragias son la TC (Tomografía Computarizada) y la resonancia magnética (RM) (Franco Martinez, 2011). Estas ayudan a determinar su tamaño, localización y crecimiento del hematoma, sin embargo, la TC es la técnica más utilizada debido a su efectividad ya que presenta una sensibilidad y especificidad cercanas al 100%. Como se expuso en el estudio de Brott (T. Brott, 1997), la TC es capaz de documentar la evolución del sangrado y permite controlar el tamaño y forma de la hemorragia. Otra gran ventaja de la TC es que requiere un menor tiempo de exploración y posee una mayor disponibilidad. Esto es muy importante ya que gran parte de los pacientes presentan un incremento del volumen de la hemorragia considerable a las tres horas del inicio de los síntomas.

En este contexto y con el fin de agilizar el diagnóstico de la manera más precisa posible, a lo largo de estos últimos años se han realizado innumerables esfuerzos usando Aprendizaje Automático. Gracias al Deep Learning, se han conseguido innumerables avances, en las que se proponen diversas técnicas para tratar de hacer frente a la detección de ICH a través de imágenes médicas, principalmente o, al menos las que más éxito han tenido, bajo el paraguas de las redes neuronales profundas: (Darias Plasencia, 2019), (Ye, 2019), (Arbabshirani, 2018). La mayoría de los modelos revisados tienen suficiente capacidad para poder ser

utilizados en entornos clínicos reales debido al potencial que tienen para detectar ICH a través de imágenes médicas incluso, en algunos casos, mejor que un radiólogo.

Arbabshirani, en su trabajo (Arbabshirani, 2018), demostró un impacto positivo en la optimización del flujo de trabajo en radiología construyendo un algoritmo de inteligencia artificial que ayudaba a priorizar las listas de trabajo de dicha área en base a la criticidad del paciente con el fin de reducir el tiempo de atención en un 96%. Además, conseguía identificar ICH muy sutiles pasadas por alto por algunos radiólogos. Para ello, empleó una red neuronal convolucional profunda entrenándola con más de 2 millones de tomografías computarizadas de diferentes tamaños debido a las diferencias de tamaño de los cráneos, por lo que se tuvo que hacer un redimensionado o resizing a 256x256. Además, se emplearon técnicas de Windowing aplicando un filtro de cerebro para identificar las hemorragias (centrada en 40HU y ancho de 80HU). Por último, para aumentar los datos de entrenamiento y conseguir un conjunto de datos más equilibrado, se aplicó DA sobre todo el conjunto de entrenamiento donde se realizaron aproximadamente entre 20 y 80 observaciones aumentadas de cada estudio negativo y estudio positivo, respectivamente. Se aumentó aplicando una traslación horizontal y vertical aleatoria (±0-20 píxeles), rotación (±0-15°) y volteado (horizontal).

Por otro lado, el trabajo de Darías (Darias Plasencia, 2019), publicado recientemente, a pesar de no ser un estudio enfocado en la dirección del diagnóstico de ICH directamente, ha trabajado en una línea muy similar tratando de detectar Alzheimer a través de imágenes médicas de RMI a través del uso de técnicas de Deep Learning. Este trabajo se basó principalmente hacer una comparativa del resultado proporcionado por dos de las redes neuronales profundas más utilizadas hoy día InceptionV3 y ResNet50 para identificar la enfermedad citada. Darías, debido a que el data set utilizado no era muy grande, se encargó de utilizar técnicas de fine-tuning sobre ambas redes con el fin de aprovechar el preentrenamiento de ambas y conseguir mejor resultado. Como técnicas de preprocesado utilizó técnicas de redimensionamiento y normalización de imágenes con el fin conseguir un entrenamiento óptimo de la red. Este estudio ha sido muy importante en el desarrollo de esta investigación ya que plantea, en su apartado de líneas futuras, una serie de recomendaciones sobre qué caminos se deben seguir a fin de seguir avanzando en la investigación proponiendo hacer uso de técnicas como Windowing o Data Augmentation.

Por otro lado, el trabajo de H. Ye (Ye, 2019) publicado recientemente hace menos de un año, es uno de los trabajos de más relevancia en el mundo de la detección hemorragias intracraneales por lo que también se ha tenido muy en consideración. En este trabajo, se evalúa la capacidad de predicción de ICH sobre tomografías computarizadas y sus subtipos usando una combinación de red neuronal convolucional (CNN) con una red neuronal

recurrente (RNN) demostrando que incluso la capacidad de precisión media era incluso mejor que la de algunos radiólogos junior. Las técnicas de preprocesamiento empleadas en este trabajo para poder entrenar al modelo son:

- Resizing: Reducción del tamaño de las TC de 512x512 a 256x256: con el fin de poder reducir el uso de memoria de la GPU.

- Windowing usando 3 ventanas diferentes de intensidad: con el fin de tener en cuenta el alto rango de intensidad dinámica que poseen las imágenes preservando, al mismo tiempo, los detalles de los diferentes objetos de interesan: (huesos, tejidos, blandos, vasos sanguíneos, …)

- Data Augmentation: con el fin de reducir overfitting y equilibrar el desbalanceo de datos aplicando técnicas de rotación (0-180 grados), reescalado (90%-110%), deslizamiento (10%), recortes (10%) y volteado vertical y horizontal.

En cuanto a la redistribución de las clases dentro del conjunto de datos usado, emplearon un data set de 76.621 tomografías computarizadas donde se podía apreciar que un claro desbalanceo de los datos:

- El porcentaje de casos Negativos es del 69%
- El porcentaje de casos Positivos es del 31%
- El porcentaje de casos Positivos con Hemorragia de tipo EPI es del 1% del conjunto del dataset.
- El porcentaje de casos Positivos con Hemorragia de tipo INP es del 11% del conjunto del dataset.
- El porcentaje de casos Positivos con Hemorragia de tipo INV es del 6% del conjunto del dataset.
- El porcentaje de casos Positivos con Hemorragia de tipo SAH: 9% del conjunto del dataset.
- El porcentaje de casos Positivos con Hemorragia de tipo SDA: 3% del conjunto del dataset.

Por último, en cuanto a los resultados de accuracy y F1 Score presentados en este trabajo para cada uno de los subtipos son los presentados en la Tabla 1:

Tabla 1
Resultados Modelo (Ye, 2019)

	Accuracy	F1 Score
EPI	0.96	0.72
INP	0.90	0.93
INV	0.91	0.87
SAH	0.83	0.78
SDA	0.94	0.84

Tabla 1 Resultados de accuracy y F1 Score del modelo con mejor resultado del trabajo de H. Ye (Ye, 2019). Estos datos serán utilizados como referencia.

2.4 Conclusiones

Tras el análisis de la literatura actual en base a los conceptos clínicos técnicos y del estado del arte expuestos en todo este capítulo, se puede concluir que existen técnicas cuyo aporte es claro e indiscutible, pero, sin embargo, existen otras que no queda del todo demostrada su efectividad por lo que no se puede concluir completamente qué técnicas de preprocesamiento o métodos de entrenamiento son los que mejor resultado acaba proporcionando.

En base a lo explicado en este capítulo, el redimensionamiento o resizing puede catalogarse como una técnica cuyo aporte es indiscutible, se identifica que el uso generalizado de esta técnica en todas las literaturas analizadas, (Darias Plasencia, 2019), (Ye, 2019), (Arbabshirani, 2018) estableciendo las imágenes disponibles con unas dimensiones fijas y comunes. Además, otra de las técnicas donde el aporte también es evidente es la técnica de la normalización o whitening con el fin de entrenar los modelos de una manera más eficiente y rápida para conseguir una mejor convergencia de la red. En cuanto a los modelos utilizados, también ha tenido un uso generalizado las redes neuronales, más en concreto, el uso de redes neuronales previamente entrenadas con ImageNet y hacer el uso de técnicas como Transfer Learning y fine-tuning, (Darias Plasencia, 2019), (Ye, 2019), (Esteva, 2017), (Gulshan, 2016). Por lo tanto, este trabajo se centra en el uso de esta técnica.

Por otro lado, en cuanto a los métodos de balanceo de clases no parece que sea muy generalizado el uso de técnicas a priori que permitan equilibrar el conjunto de datos para el entrenamiento o, al menos, no se menciona sí se ha realizado o no. La única que técnica más extendida es DA pero es complicado determinar el grado de aportación que aplica esta técnica a los resultados de los modelos. Por lo tanto, este punto es objeto de estudio en este trabajo.

En relación con otra de las técnicas que, en base a la literatura analizada, parece tener una gran importancia y un uso extendido es la técnica de Ventanas de TC o Windowing. En los trabajos de Ye y Arbasbshirani (Ye, 2019), (Arbabshirani, 2018), es utilizada esta técnica aplicando diferentes ventanas con el fin de extraer las características más importantes de las TC.

Finalmente, en base a las recomendaciones de líneas futuras que hizo Darías (Darias Plasencia, 2019), se ha establecido también como objeto de estudio el uso de técnicas de preprocesamiento como Windowing y métodos de balanceo de datos como Data Augmentation con el fin de comprobar cómo de efectivo es el uso de las mismas. Además, se establecerá como punto de referencia los resultados ofrecidos por el trabajo de Ye (Ye, 2019) que, en base a la literatura inspeccionada, ha sido el estudio que mejor resultado ha proporcionado hasta la fecha.

Capítulo 3. Objetivos y Metodología de Trabajo

En este tercer capítulo, se exponen los objetivos y metodología de trabajo que se han llevado a cabo para elaborar esta tesis.

3.1 Objetivos Generales

El objetivo general de este estudio es hacer recomendaciones sobre el uso de diferentes técnicas que ayuden al entrenamiento de una red neuronal con el fin de mejorar la detección de hemorragias intracraneales sobre tomografías computarizadas realizando una comparación experimental de las técnicas expuestas.

3.2 Objetivos Específicos

Los objetivos específicos han sido los siguientes:

1. Revisión sistemática de la literatura actual con relación a la detección de hemorragias intracraneales a nivel clínico describiendo el contexto de aplicación de este y aportando un resumen del conocimiento que existe en este campo.

2. Identificar los modelos, técnicas de preprocesamiento y métodos de balanceo de datos para el entrenamiento de una red neuronal que mejor resultado han proporcionado para la detección de hemorragias intracraneales, a través de una revisión sistemática de la literatura actual con el fin de elaborar un estudio del estado del arte.

3. En base a las carencias identificadas en los objetivos anteriores, definir un banco de pruebas e implementar los modelos considerados que pueden ayudar a mejorar la capacidad de predicción de un sistema artificial de detección de ICH.

4. Entrenar los modelos de cada una de las pruebas definidos en el objetivo específico 3 analizando sus resultados y la capacidad de predicción de cada modelo diferentes tipos de hemorragia.

5. Comparar los resultados obtenidos por los diferentes modelos entrenados en el objetivo específico 4, analizarlos y emitir las recomendaciones oportunas.

3.3 Metodología de trabajo

El proceso de investigación en esta tesis se dividió en dos fases principales: la fase exploratoria y la fase comparativa experimental.

La fase exploratoria se llevó a cabo para identificar aquellas técnicas que más se usan en el diagnóstico automatizado de las ICH. En particular se realizaron las siguientes actividades en esta fase:

- Elaboración de una revisión sistemática de la literatura actual que permitió una mejor comprensión de las ICH y su diagnóstico a nivel clínico.
- Elaboración de una revisión sistemática de literatura que permitió una mejor comprensión de la aplicación del Aprendizaje Automático en el diagnóstico de las ICH actualmente.
- Identificar los modelos, técnicas de preprocesamiento y métodos de balanceo de datos para el entrenamiento de una red neuronal que mejor resultado han proporcionado para la detección de ICH a través de una revisión de la literatura actual. Como resultado, se proporciona información sobre los modelos que serían más indicados a la hora de desplegar un sistema de diagnóstico de ICH en un entorno real, y sobre los principales métodos de preprocesamiento y técnicas balanceo de datos necesarios para ponerlos en práctica.

Por otro lado, la fase descriptiva se realizó con el objetivo de describir el funcionamiento y comportamiento de los modelos seleccionado en la fase anterior y comparar de manera experimental sus resultados. Para ello se realizaron las siguientes actividades:

- Definición de un banco de pruebas con los diferentes modelos, técnicas y métodos identificados en las conclusiones del estado del arte con el fin de preparar la comparativa.
- Implementación de los modelos usando las diferentes técnicas y métodos definidos en el banco de pruebas.
- Ajuste o balanceo de los conjuntos de datos y entrenamiento de los modelos en base a las directrices indicadas en la definición de cada una de las pruebas.
- Obtención de resultados del entrenamiento y la evaluación de cada uno de los modelos.

- Valoración final teniendo en cuenta los resultados obtenidos, analizándolos y comparando los mismos y contrastando con la información extraída durante la fase exploratoria.

- Realizar recomendaciones pensadas para profesionales de Inteligencia Artificial sin conocimientos profundos de medicina. Partiendo de las conclusiones del punto anterior, se busca proporcionar una base sobre la que partir a la hora de construir nuevos modelos de diagnóstico o la expansión de este.

Capítulo 4. Planteamiento de la Comparativa Experimental

En este capítulo, se comienza identificando el problema concreto y el objetivo para acometerlo, así como el alcance. Seguidamente, se describen los datos que se han utilizado, así como sus detalles y distribución. Más adelante, se exponen las diferentes técnicas de preprocesamiento de los datos y el modelo de red neuronal que se van a utilizar en este estudio. Posteriormente, se presenta los entornos e infraestructura usada y se presenta la estrategia de entrenamiento, así como una breve descripción esquemática de las soluciones que se proponen a estudio. Finalmente, se indican qué criterios de éxito se han considerado, así como las métricas que se han utilizado.

El objetivo de esta comparativa es averiguar cuáles de las técnicas presentadas en el capítulo 2 (preprocesamiento de ventanas de TC o Windowing, métodos de balanceo de datos y Data Augmentation), proporcionan mejor resultado para la detección de ICH y la clasificación de subtipos en TC cerebrales. Para ello, se han definido una serie de pruebas que utilizan las diferentes técnicas con el fin de tratar de comparar los resultados entre sí y comprobar cuál es la que mejor resultado proporciona. Además, se utilizan los resultados obtenidos en el trabajo de Ye (Ye, 2019) como referencia para comparar con los resultados obtenidos en estas pruebas y comprobar la efectividad de estas técnicas con otros trabajos.

Debido al reducido tiempo que se ha tenido para poder desarrollar este Trabajo de Fin de Master, se ha tenido que seleccionar con mucho cuidado cuales han sido las pruebas candidatas a realizar. Para ello, se ha buscado el apoyo en la fase exploratoria donde se ha podido extraer las técnicas de preprocesamiento y los diferentes algoritmos de entrenamiento que han protagonizado la investigación en este campo en los últimos años. Algunas de las técnicas, sobre todo aquellas que su aportación era muy evidente, han sido utilizadas en todas las pruebas. Por el contrario, el resto de las técnicas cuyo aporte necesitaba ser evaluado, se utilizan en algunas pruebas y en otras se ha descartado con el fin de comprobar su aporte y efectividad a la hora de detectar ICH y clasificar los subtipos.

4.1 Conjunto de datos

La pieza más importante en un problema de clasificación con IA son los datos. Para esto se ha utilizado un conjunto de datos proporcionado por la Sociedad Radiológica de América del Norte o, en inglés, Radiological Society of North America (RSNA) (Radiological Society of North America, 1915), en colaboración con los miembros del American Society of Neuroradiology (ASNR) (American Society of Neuroradiology, 1965) que ayudaron en el etiquetado de los datos.

La Sociedad Radiológica de América del Norte (RSNA) es una sociedad internacional de radiólogos, físicos médicos y otros profesionales de la medicina con más de 54.000 miembros en todo el mundo. Esta sociedad organiza reuniones y publica revistas todo con el fin de fomentar la educación e investigación de la neurorradiología.

Por otro lado, la ASNR es la principal organización mundial para el futuro de la neurorradiología y representa a más de 5.300 radiólogos, investigadores, intervencionistas y científicos de imagen. Esta institución organizó un cuadro de más de 60 voluntarios para etiquetar los datos.

De los datos cedidos por la RSNA, se han obtenido 674.257 tomografías computarizadas de más de 25.000 exámenes clínicos realizados a más de 17.079 pacientes. Los datos fueron proporcionados totalmente anonimizados y cada tomografía tenía asignada 6 etiquetas dónde se establecía valor ‘1’ o ‘0’ en caso de ser padecer o no la hemorragia asociada:

- ANY: paciente con algún tipo de ICH.
- EPI: paciente con un tipo de ICH Epidural.
- INP: paciente con un tipo de ICH Intraparenquimatosa.
- INV: paciente con un tipo de ICH Intraventricular.
- SAH: paciente con un tipo de ICH Subaracnoidea.
- SDA: paciente con un tipo de ICH Subdural.

En la Ilustración 15 y 16 se puede apreciar la distribución de los datos del reparto de cada una de las clases citadas, así como el total y el numero de pacientes negativos y positivos.

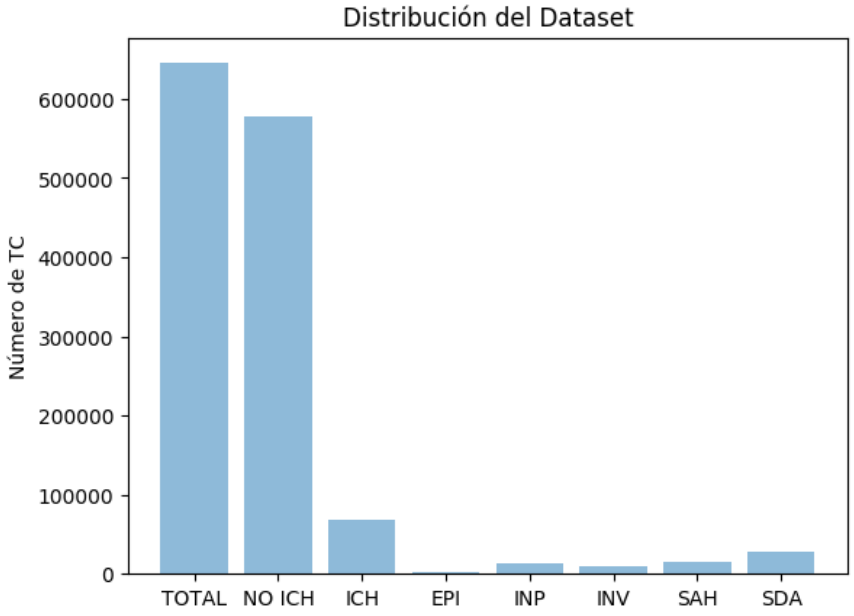

Ilustración 15 Distribución del Conjunto de Datos proporcionado por la RSNA

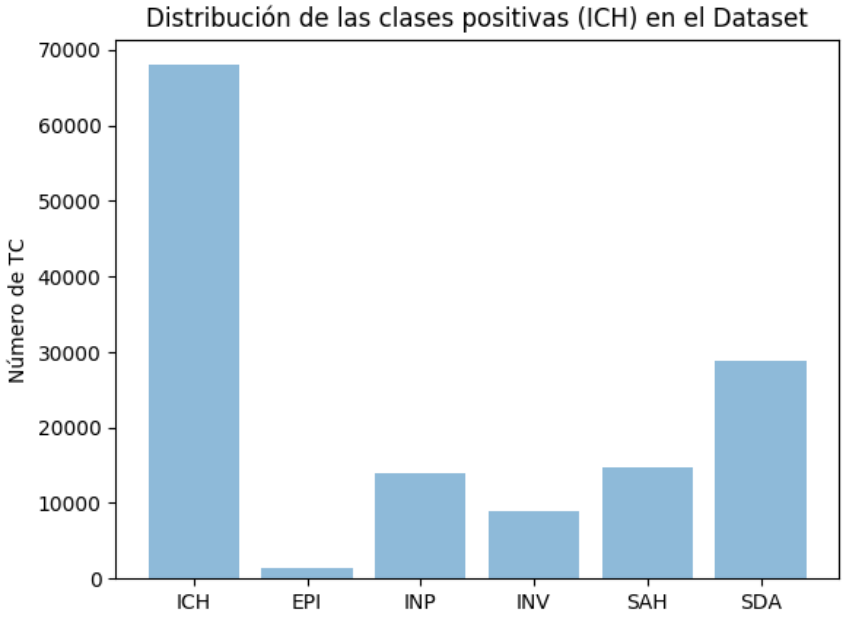

Ilustración 16 Distribución del Conjunto de Datos proporcionado por la RSNA (Clases positivas)

Como se puede apreciar en las Ilustraciones 14 y 15, el conjunto de datos está desbalanceado considerablemente. Como suele ser habitual en los conjuntos de datos médicos para la detección de enfermedades, la clase positiva ICH, es decir, tomografías que tienen algún tipo de hemorragia, únicamente tiene una representación de un 10 % respecto al total de imágenes. Por otro lado, la clase positiva EPI, es decir, tomografías que tienen el tipo de

hemorragia Epidural, es la clase con menor representación del dataset con tan solo un 0,2 % respecto al total.

Con el fin de que el reparto entre las clases positivas (Hemorragias de tipo Epidural, Intraparenquimatosa, Intraventricular, Subaracnoidea y Subdural) frente a las negativas sea más equilibrado, se ha decidido aplicar la técnica de undersampling a las tomografías de la clase negativa pasando de 577.154 imágenes o cortes a 57.983. De este modo, el reparto entre clases positivas y negativa se quedó en un 54% frente a 46%. A pesar de las recomendaciones de Mazurowskia (Maciej A. Mazurowskia, 2008) que no aconsejaba el uso de undersampling, se decidió usarla por dos motivos: el primero, por el tiempo disponible, con esta técnica se reducía el conjunto de datos en un 80 % y, segundo, porque se ha dado prioridad a las clases positivas frente a las negativas ya que en un problema de clasificación de padecimiento de ICH, como ha sido este caso, es preferible que el modelo se equivoque en el caso de que una imagen etiquetada como negativa se prediga como positiva en vez de lo contrario. Sería más preocupante si un paciente con ICH se le clasificase como un paciente sano. Por lo tanto, se ha considerado que hacer esta reducción no debería impactar de una manera significativa en la comparativa de resultados que es objetivo de este estudio. Por otro lado, siguiendo las recomendaciones de (Maciej A. Mazurowskia, 2008) la técnica de oversampling fue descartada por elevar el riesgo de overfitting.

Finalmente, la distribución final que se ha quedado en el conjunto de datos se presenta a continuación en la Tabla 2. Además, se aprovecha para indicar la distribución entre conjunto de entrenamiento, validación y test:

Tabla 2
Distribución del conjunto de datos

	Train	**Validación**	**Test**	**Total**	
NEG	40.401	5.774	11.808	**57.983**	*46%*
EPI	1.047	149	301	**1.497**	*1%*
INP	9.752	1.380	2.800	**13.932**	*12%*
INV	6.239	883	1.792	**8.914**	*7%*
SAH	10.301	1.458	2.958	**14.717**	*11%*
SDA	20.236	2.863	5.810	**28.909**	*23%*
Total	**87.976**	**12.507**	**25.469**	**125.952**	
	70%	*10%*	*20%*		

Tabla 2 Distribución del conjunto de datos proporcionado por la RSNA (Radiological Society of North America, 1915) a utilizar en este estudio.

4.2 Pre-Entrenamiento

Antes de entrenar el modelo, se requiere un cierto preprocesamiento antes de que los datos estén listos para ser consumidos por la red neuronal. Se detallan a continuación:

- Resizing

  Siguiendo los pasos del trabajo de Ye (Ye, 2019), se ha decidido realizar un redimensionamiento de las imágenes del dataset con el fin de establecer un tamaño común de las imágenes médicas para todos los modelos y tratar de agilizar el entrenamiento de estos ya que permite reducir bastante el uso de memoria de la GPU. Tras diversas pruebas, se tomó la decisión de utilizar una dimensión de 256x256, un valor de dimensión bastante adecuado ya que era capaz de mantener en su totalidad los detalles de la imagen sin tener que utilizar unos tamaños de imagen tan grande como 512x512.

  Tal y como se concluyó en el capítulo 2, en base a los trabajos estudiados (Darias Plasencia, 2019), (Ye, 2019), (Arbabshirani, 2018) quedó demostrado la aportación del redimensionamiento es indiscutible por lo que se ha mantenido este método en todas las pruebas que realizadas en este estudio.

- Windowing

  Como se ha explicado en el punto 2, en base a los trabajos estudiados (Ye, 2019), (Arbabshirani, 2018) el objetivo de esta técnica es ser capaces de identificar diferentes tipos de tejidos dentro de la TC. En este trabajo se comprueba cuanto de efectiva es esta técnica por lo que en aquellos modelos donde se aplique Windowing, se realiza lo siguiente:

  Las tomografías del dataset han sido proporcionadas por la RSNA en el formato DICOM. DICOM es el estándar que más se utiliza para almacenaje e intercambio de imágenes médicas obtenidas por TC. Este formato, además de contener la imagen médica en sí, contiene muchos metadatos tales como: el tamaño de los píxeles, la longitud de un píxel en cada dimensión del mundo real, …

  La unidad de medida en las tomografías es la Unidad Hounsfield (HU), que es una medida de la radiodensidad. Los escáneres TC están cuidadosamente calibrados para medir esto con precisión. Desafortunadamente, los valores devueltos no están en esta

unidad. Por lo tanto, es necesario hacer unos calculos para poder trabajar en esta unidad.

Algunos escáneres tienen límites de escaneo cilíndricos, pero la imagen de salida es cuadrada. Los píxeles que caen fuera de estos límites obtienen el valor fijo de -1024, por lo que el primer paso será fijar estos valores a -1024, valor que corresponde al aire como se ha explicado en el apartado 2 (Franco Martinez, 2011).

Seguidamente, se multiplica el valor de cada píxel por la pendiente de reajuste y se suma con la intercepción. Ambos valores se pueden extraer de los metadatos del fichero DICOM a través de los campos RescaleSlope y RescaleIntercept respectivamente.

Una vez se ha convertido la imagen a HU, se procede a aplicar las respectivas ventanas con el fin de extraer únicamente los tejidos más relevantes para la identificación de ICH ya que cada tipo de tejido está asociado a un nivel de HU. Los tres tipos de ventanas elegidas son:

- Brain Window o ventana del cerebro, cuyos valores de ventana son centro: 40 HU y ancho: 80 HU es capaz de extraer las estructuras propias del cerebro.
- Subdural Window o ventana subdural, cuyos valores de ventana son centro: 80 HU y ancho: 200 HU con el fin de hacer más visible la ICH. Ayuda en la detección de los delgados hematomas subdurales agudos, los más difíciles de detectar debido a su
- Soft Window o ventana de tejidos blandos, cuyos valores de ventana son centro: 40 HU y ancho: 380 HU con el fin de extraer los nervios, vasos sanguíneos y diferentes membranas que puedan ayudar a la identificación de la hemorragia.

Después de aplicar cada ventana, se combinan en una sola imagen de 3 canales. Cada canal de la imagen corresponde a una de las ventanas generadas teniendo como resultado una imagen de 256x256x3 que será la entrada a la red neuronal que se explicará con detalle más adelante.

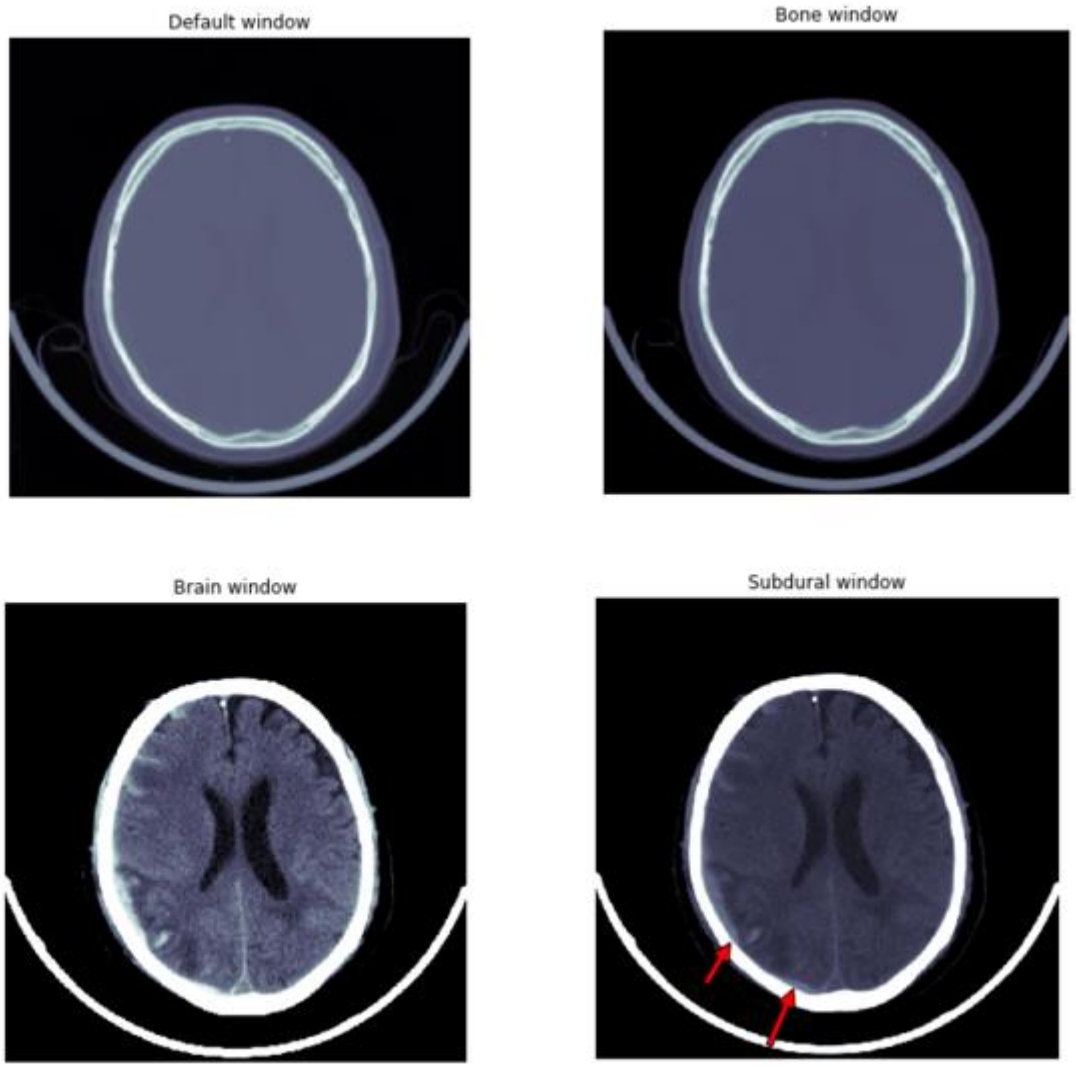

Ilustración 17 Visualización del mismo corte axial de una tomografía computarizada aplicando diferentes ventanas. La selección de una ventana u otra permite poder visualizar los detalles dependiendo del tipo de tejido (radiopedia.org)

Como ejemplos, se puede apreciar en la Ilustración 17, en la imagen superior izquierda se tiene la visualización de la imagen sin la aplicación de ninguna ventana. En la inferior izquierda, se ha aplicado la ventana del cerebro donde se puede apreciar con mayor detalle la fisionomía del cerebro, ventrículos, la fisura interhemisférica, pliegues… En la superior derecha, se ha aplicado una ventana de huesos con el fin de destacar el cráneo. Y, por último, en la inferior derecha, se ha aplicado la ventana subdural que permite visualizar la ICH. Nótese la importancia de la aplicación de esta técnica con un tipo de ventana u otro para poder destacar e identificar este tipo de hemorragia.

Finalmente, para aquellos modelos dónde no se utilice la técnica de Windowing se obtiene el array de pixeles directamente del fichero DICOM y se expanden los mismos valores a los otros 2 canales con el fin de establecer el tamaño de entrada que exige la red neuronal 256x256x3.

- Whitening o Normalización

  En base al trabajo de Ye y Pal & Sudeep, (Ye, 2019), (Sudeep, 2017), se ha decido aplicar normalización en todos los modelos ya que, tal y como se indicó en el capítulo 2, su aportación es totalmente evidente. Primero se obtiene la imagen de 256x256x3 fruto de la salida de la aplicación de Windowing o Sin Windowing y se reduce el valor de la intensidad de cada pixel en valores reales entre 0 y 1 resultante de restar la media y dividirlo por la desviación típica.

4.3 Arquitectura del Modelo

Como se ha comentado previamente en la fase exploratoria, el uso de redes neuronales para la resolución de este problema de clasificación es una tendencia muy adoptada en los últimos años para este tipo de problemas de clasificación con imágenes médicas (Darias Plasencia, 2019), (Ye, 2019), (Arbabshirani, 2018), (Esteva, 2017), (Ding, 2018). Debido a que el tiempo disponible para el desarrollo de este estudio ha sido bastante reducido, se ha decido evaluar qué técnicas de preprocesado y algoritmos de entrenamiento tienen mejor resultado usando un tipo de red neuronal base. Por esto y por evitar añadir más variables al juego que puedan confundir a la hora de determinar qué algoritmo o técnica es mejor, se ha decido utilizar sólo una red neuronal.

Tras diversas pruebas para poder elegir la red que mejor se ajuste, se tomó la decisión de utilizar una de las redes que proporciona Keras dentro de su librería: la red neuronal ResNet 50. Esta red es una de las más utilizadas y que mejor resultado da para este tipo de problemas de clasificación. Este tipo de redes han sido preentrenadas con un set de imágenes de ImageNet.

Este tipo de redes están diseñadas para trabajar con imágenes de 3 canales RGB, de ahí la necesidad en el preprocesado de que por cada TC era necesario la expansión a 3 canales. Por lo tanto, el input para la red neuronal ha sido de 256x256x3.

Para todas las pruebas, se ha aplicado la técnica de Transfer Learning haciendo fine-tuning sobre la ResNet 50, es decir, que únicamente se aprovechan los pesos preentrenados con ImageNet de las primeras capas, cuyos parámetros no se alteran quedándose fijos durante todo el entrenamiento. En cuanto a las últimas capas preentrenadas de la red se desechan y se sustituyen por tres capas:

- 2 capas de 64 neuronas Fully Connected
- 1 capa Softmax de 5 neuronas de salida que representan las 5 posibles clases (EPI, INP, INV, SHA y SDA).

La Ilustración 18 muestra un ejemplo gráfico de la arquitectura común utilizada en los modelos de este trabajo.

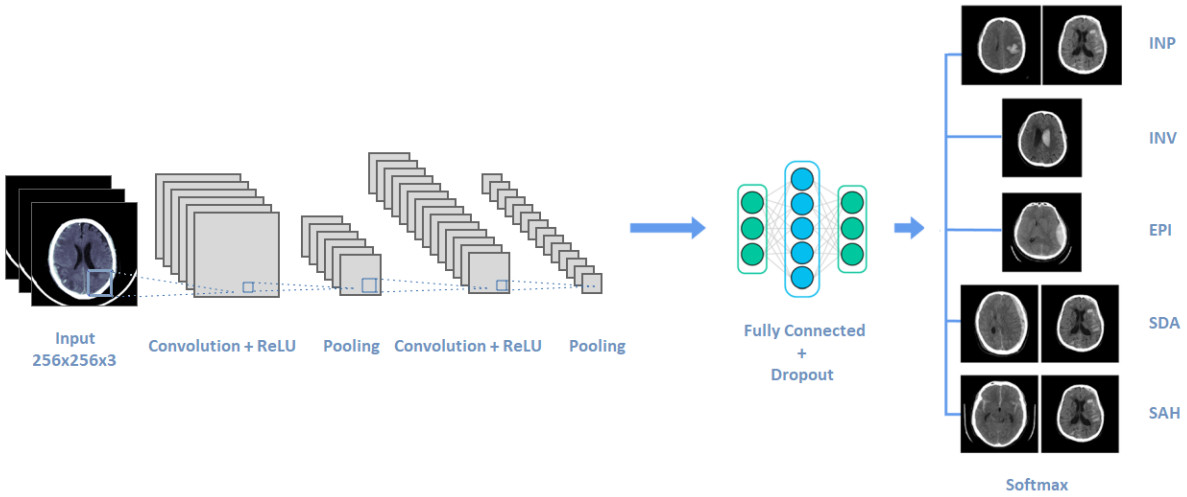

Ilustración 18 Arquitectura común para los modelos en estudio

Los parámetros de estas capas se ajustarán durante el entrenamiento de la red. Además, se ha añadido una regularización dropout del 30% con el fin de evitar el posible overfitting.

El motivo principal del uso de fine-tuning es el ahorro en recursos de computación y tiempo de entrenamiento. La ResNet 50 previamente ha aprendido a extraer las características universales necesarias por lo que con esta técnica se transfiere ese aprendizaje a nuestra nueva red. Lo único que quedaría es ajustarla a nuestro gusto entrenando las últimas capas con el conjunto de imágenes de la RSNA.

Por último, tras una numerosas de pruebas se decidió utilizar el optimizador avanzado RMSProp con una tasa de aprendizaje o learning rate de 10^{-4} y la entropía cruzada categórica (categorical cross-entropy) como función de coste o pérdida.

4.4 Infraestructura, Estrategia de Entrenamiento y Balanceo de Clases

Para empezar a desarrollar y entrenar los modelos se ha buscado apoyo en las diferentes plataformas que existen en la nube. Estas plataformas reducen el tiempo de preparación de entornos ya que proporcionan gran cantidad de software previamente instalado y configurado. Además, de esta manera el despliegue de la infraestructura es menos costosa que un entorno on-premise ya que todos los recursos hardware se pagan en función de su uso.

Una de las plataformas más usadas hoy día y totalmente gratuita es la de Google Colaboratory, sin embargo, ésta se queda pequeña debido a que se tiene que utilizar más de 125.000 imágenes, el espacio en disco y los 12GB de RAM disponibles que proporciona la plataforma se quedan muy cortos, por lo que se debe buscar una forma alternativa de alimentar los modelos.

La plataforma elegida ha sido “Google Cloud Platform”. Principalmente, se ha utilizado una instancia con el software preinstalado de TensorFlow 2.1 (Google, 2020) con soporte de Keras (Chollet F. , 2015). Dentro de esta instancia se dispone del entorno de desarrollo Jupyter Lab (Jupyter, 2015) que ha permitido la codificación de todos los modelos en Python. En cuanto al hardware de la instancia se aprovisionó un entorno con 8 vCPUs con 30 GB de RAM y además 4 GPUs Tesla K80.

En cuanto al almacenamiento en disco se ha requerido aprovisionar en este entorno 1500 GB de SSD de memoria persistente. Con el fin de agilizar el entrenamiento se ha utilizado memoria SSD ultra rápida siendo capaz de llegar a una lectura de 720 MB/s.

Una de las razones utilizar tal cantidad de memoria en disco es debido a que, como se ha comentado previamente, es necesario entrenar modelos con un set de imágenes bastante grande por lo que se necesita mucho espacio en memoria. Si se trata de entrenar el modelo directamente es probable que antes de terminar la primera época (epoch en inglés) salte un aviso de que no hay espacio en memoria RAM. Por ello, se ha elegido utilizar los Data Generators nativos de Python (Amidi, 2018) que son una opción muy utilizada en Keras. Éstos permiten leer la información directamente desde el disco duro en lugar de desde la memoria RAM. Es un proceso más lento, pero evita la interrupción del entrenamiento por falta de memoria RAM. Además, estos Data Generators dan la posibilidad de la realización del pre procesado (Resizing, Windowing/No Windowing y Normalization) y la aplicación de Data Augmentation durante el entrenamiento batch a batch por lo que ayudan a simplificar mucho más el código.

Por otro lado, el tamaño de los batches elegidos ha sido de 32 elementos tras diversas pruebas realizadas. En cuanto a la estrategia de entrenamiento a seguir se ha tenido muy en

consideración el desbalanceo de los datos por lo que ha sido necesario seleccionar varias alternativas de tal manera que puedan ser comparados sus resultados a posteriori.

Como se ha comentado previamente la distribución de los datos, el balanceo entre clases positivas y negativas está relativamente balanceado alrededor del 54% y el 46% respectivamente. Sin embargo, el objetivo de este trabajo no es detectar únicamente si el paciente en esa TC tiene o no ICH si no que, además, es necesario averiguar el tipo de ICH que padece. Es aquí dónde radica el problema ya que tenemos las clases de los tipos tenemos las clases resultantes muy desbalanceadas dentro del dataset. Se han propuesto 2 estrategias para poder comprobar la efectividad entre ellas:

- Fijo y desequilibrado: No alterar el contenido del conjunto de entrenamiento en cada época. De esta manera, cada época procesa las 87.976 TC tal y como se aprecia en la tabla 2. Con esta estrategia mantendremos un claro desbalanceo de los datos.

- Dinámico y equilibrado: Alterar el contenido del conjunto de entrenamiento en cada época de tal manera que el número de datos procesados de cada clase en cada época sea el mismo. De esta manera, cada época procesa las 6.000 TC extrayendo de manera aleatoria 1.000 TC de cada clase. Con esta estrategia evitaremos que la red se entrene en cada época con más datos de una clase que de otra.

Por otro lado, otra de las alternativas que tratan de resolver el problema del desequilibrio de la representación de las clases es la utilización de la técnica de Data Augmentation (DA) en el entrenamiento de los datos. Está técnica era una de las líneas futuras de recomendación que hizo Darías en su trabajo (Darias Plasencia, 2019). Con esta técnica lo que se busca es tener la capacidad de expandir el volumen de datos de nuestro dataset realizando una serie de transformaciones en los datos con el fin para enriquecer el entrenamiento del modelo con nuevas observaciones a partir de las que ya existen. Se han propuesto 3 estrategias para poder comprobar la efectividad entre ellas:

- No usar DA. La red neuronal recibirá las observaciones tal y como son, sin establecer ninguna transformación en las imágenes médicas.

- Uso de DA aplicado a todas las imágenes del dataset. La red neuronal obtendrá cada una de las observaciones del dataset con una transformación aplicada de tal manera que aumenta la variedad entre las mismas.

- Uso de DA aplicado a únicamente a las imágenes del dataset etiquetadas como tipo de ICH Epidural. La red neuronal obtendrá cada una de las observaciones del dataset

tal y como es sin aplicar ningún tipo de transformación salvo en las de tipo Epidural las cuales se aplicará una transformación de tal manera que aumenta la variedad entre las mismas consiguiendo elevar la representación de esta clase frente al resto.

Con el fin de afectar lo menos posible al objetivo principal de este trabajo, en aquellos elementos donde se aplica DA, se ha decidido utilizar la misma configuración, es decir, siempre que se aplique DA, ya sea en una sola clase o en todo el conjunto de datos, se aplican las siguientes técnicas:

- Volteado horizontal a un 25%.
- Volteado vertical a un 10%.
- Recorte por cada lado de la imagen por un valor de píxeles aleatorios muestreados uniformemente en el intervalo discreto entre 0 y 25.

Además, dentro del abanico de posibilidades que proporciona DA se ha preferido elegir una configuración sutil, no muy sobrecargada o agresiva, con el fin de identificar si el uso de esta técnica es apropiado o no. De hecho, esta configuración sólo se aplica al 25% de las observaciones de cada batch.

Las diferentes estrategias de DA, así como los diferentes algoritmos de entrenamiento que se acaban de presentar, son totalmente compatibles para ser combinados por lo que en este trabajo se ha definido una serie de pruebas que ayudan a comparar de mejor manera cada una de estas técnicas. A continuación, en la Tabla 3 se exponen cada una de las pruebas que se han decidido realizar:

Tabla 3
Listado de pruebas

Prueba	Windowing	Data Augmentation	Estr. de Entrenamiento
1	No	Sólo en Epidural	Dinámico y equilibrado
2	No	En todo	Dinámico y equilibrado
4	Sí	En todo	Fijo y desequilibrado
5	Sí	No	Fijo y desequilibrado
6-2B	Sí	En todo	Dinámico y equilibrado
7-1B	Sí	Sólo en Epidural	Dinámico y equilibrado
8	SI	No	Dinámico y equilibrado

Tabla 3 Listado de pruebas a ejecutar en este trabajo. Se identifican las características más importantes y diferenciadoras entre cada una de ellas.

Con el fin de no querer extender este estudio más de lo necesario y centrarse en los objetivos más importantes, se han omitido algunas pruebas del ámbito de este estudio. Se han realizado más pruebas (3, 9 y 10) pero han quedado fuera del ámbito de este trabajo por lo que no se ha hecho mención.

4.5 Criterios de Éxito y Métricas

Como se ha comentado previamente, el objetivo es comparar los resultados de cada una de estas pruebas con el fin de ver qué combinación aporta una mayor predicción de las ICH y sus subtipos.

Para ello, se utilizan principalmente dos métricas que aportan cómo de bueno ha sido la predicción de cada una de las clases en cada modelo o prueba:

- Exactitud o accuracy.
- F1 Score.

El primero proporciona la información de cómo de preciso es el algoritmo a la hora de identificar las diferentes clases. El segundo es la media armónica de las métricas de precisión y recall, por lo tanto, se utiliza ésta como un resumen de ambas.

La razón de utilizar más métricas aparte de la exactitud o accuracy es debido a que la información que puede proporcionar la exactitud puede llevar a engaño. Como se ha comentado con anterioridad, los conjuntos de datos médicos suelen tener muy poca representación de la clase positiva por lo que si se tiene un conjunto de test donde sólo hay 1% de clases positivas y el resto son negativas, si el modelo predijese como todas las observaciones como negativas tendría un 99% de exactitud y, sin embargo, no estará detectando lo que realmente se quiere: detectar la enfermedad. La precision indica la proporción de aquellos clasificados por nuestro modelo como positivos, estaban etiquetados realmente en nuestro conjunto de entrenamiento como tal. Por otro lado, la métrica de recall indica la proporción de aquellos que estaban etiquetados como positivos en nuestro conjunto de test si han sido clasificados como tal. Por ello, se utiliza la métrica F1 Score ya que contiene la información de precision y recall en una sola calculando su media armónica. A continuación, se muestran las fórmulas de cada una de las métricas comentadas:

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

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

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

$$F1\ Score = \frac{2\ (\ Precision \cdot Recall)}{(\ Precision + Recall)}$$

Todas estas métricas se tomarán a cada una de las clases con el fin de detectar cómo de bueno es el modelo para clasificar las observaciones en cada una de ellas. Por lo tanto, con el fin de obtener una comparación lo más real e igualitaria posible entre todos los modelos a probar, se han establecido varios subconjuntos de test donde se escogerá un número concreto de observaciones uniforme para cada las clases. Es decir, la clase con menor representación vuelve a ser epidural con 301 observaciones en el conjunto de test por lo que se establecerán los siguientes subconjuntos de test:

- Subconjunto de EPI: 300 observaciones.
- Subconjunto de INP: 300 observaciones.
- Subconjunto de INV: 300 observaciones.
- Subconjunto de SHA: 300 observaciones.
- Subconjunto de SDA: 300 observaciones.
- Subconjunto de Total: 1500 observaciones, la suma de cada una de las anteriores.

Por lo tanto, las métricas de accuracy y F1 Score han sido tomadas en cada uno de estos subconjuntos de test y se han ido chequeando al final de cada época a medida que el entrenamiento fue evolucionando con el fin identificar si el modelo entra o no en overfitting.

Por último, los resultados se contrastan con el trabajo de referencia de H. Ye (Ye, 2019)con el fin de comprobar cómo de bueno ha sido el entreno del modelo respecto al de H. Ye y comprobar la aportación realizada a la investigación.

Capítulo 5. Desarrollo de la Comparativa

Todas las pruebas presentadas en el capítulo anterior y que se detallan a continuación, comparten ciertas configuraciones y técnicas comunes tal y como se expresó en dicho apartado. Estas son: el uso de redimensionado a 256x256x3 y su posterior normalización, uso de Transfer Learning sobre la ResNet 50 previamente entrenada con ImageNet donde sólo se entrenan las Fully Connected Layers del final, función de optimización: RMSProp con Learning Rate 10-4 y su función de coste asignada Categorical Cross Entropy.

A continuación, se detalla el desarrollo de cada una de las pruebas, así como sus objetivos y resultados.

5.1 Pruebas 1 y 2

Las dos primeras pruebas han sido planteadas con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento frente a cuando se aplica únicamente a la clase con menor representación (Epidural) sobre un escenario dónde no se aplica la técnica de Windowing como preprocesamiento. Ambas pruebas, además comparten la misma estrategia de entrenamiento a la hora de distribuir las clases de una manera dinámica y equilibrada en cada una de las épocas del training.

En la primera prueba, se ha entrenado el modelo con 1000 elementos de cada clase en cada época, haciendo una selección aleatoria de los 1000 primeros elementos de cada clase. La clase Epidural sólo dispone de 1047 elementos en todo el conjunto de entrenamiento, esto significa que en la siguiente la siguiente época se volvería a procesar al menos 953 observaciones repetidas de la clase Epidural cuando para otras clases lo más probable es que procesen nuevas observaciones. Por lo tanto, por la sospecha que existe de que el modelo pueda memorizar los elementos de esta clase llegando a no generalizar por overfitting, se decide aplicar DA sólo para esta clase.

Por otro lado, en la segunda prueba, se ha entrenado de igual manera el modelo con 1000 elementos de cada clase en cada época, haciendo una selección aleatoria de los 1000 primeros elementos de cada clase. En este caso, pensando que a medida que transcurriera el entrenamiento podría producirse un agravio en las clases que no tienen aplicado DA, se decide comprobar, qué resultado se obtiene en el caso de que en todas las clases se aplique la misma técnica de DA.

El entrenamiento de ambos modelos fue similar. Se estuvieron entrenando hasta que ver que la red no proporcionaba mejores resultados. El entrenamiento llegó a las 200 épocas lo cual supuso 24 horas por cada modelo en los entornos previamente explicados en el capítulo 4.

Los resultados de ambas pruebas fueron bastante pobres. Hubo síntomas de overfitting sin tendencia de mejora.

- En la primera prueba:
    - Durante el entrenamiento, en los resultados de validación, tanto el accuracy como el loss se mantuvieron. El accuracy se quedó en torno al 72% durante prácticamente todo el entrenamiento y sin signos de mejora. Por otro lado, el los, se mantuvo con valores entre 3.1 – 4.8 aproximadamente durante prácticamente todo el entrenamiento y sin signos de mejora.
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos fue:
        - Subconjunto de Clases Negativas: pasó de tener una accuracy del 99% en las primeras épocas a un 66% sin signos de mejora.
        - Subconjunto de Clases EPI: el accuracy fue oscilando entre 66% - 74% pero no consiguió salir de ahí.
        - Subconjunto de Clases INP: presentó indicios de mejora entre las épocas 20 - 60 llegando incluso a un 80%, sin embargo, después se mantuvo en un 66%.
        - Subconjunto de Clases INV: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
        - Subconjunto de Clases SAH: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
        - Subconjunto de Clases SDA: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
        - Subconjunto Total: se mantuvo en un 72% de accuracy durante todo el entrenamiento. F1 Score sin mejorar en toda su ejecución con valores de 0.05%.

- En la segunda prueba:
    - Durante el entrenamiento, en los resultados de validación, tanto el accuracy como el loss se mantuvieron. El accuracy se quedó en torno al 72% durante prácticamente todo el entrenamiento y sin signos de mejora. Por otro lado, el

los, se mantuvo con valores entre 3.1 – 4.8 aproximadamente durante prácticamente todo el entrenamiento y sin signos de mejora.

- En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos fue:
  - Subconjunto de Clases Negativas: pasó de tener una accuracy del 99% en las primeras épocas a un 66% sin signos de mejora.
  - Subconjunto de Clases EPI: el accuracy fue oscilando entre 66% - 74% pero no consiguió salir de ahí.
  - Subconjunto de Clases INP: presentó indicios de mejora entre las épocas 20 - 60 llegando incluso a un 80%, sin embargo, después se mantuvo en un 66%.
  - Subconjunto de Clases INV: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
  - Subconjunto de Clases SAH: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
  - Subconjunto de Clases SDA: se mantuvo en un 66% de accuracy durante todo el entrenamiento.
  - Subconjunto Total: se mantuvo en un 72% de accuracy durante todo el entrenamiento. F1 Score sin mejorar en toda su ejecución con valores de 0.03%.

Nota: Debido al pobre resultado de estas dos pruebas no se tomó la decisión de obtener los datos de F1 Score para cada uno de los subconjuntos de test de los subtipos de ICH. Otro condicionante fue que durante la fase de evaluación de estos modelos sólo se recogía esta métrica para el subconjunto Total con el fin de tener una primera idea rápida de cómo de bueno estaba siendo el modelo. No obstante, a partir de las siguientes, este dato se recogió para todos los subconjuntos.

5.2 Pruebas 4 y 5

Esta nueva pareja de pruebas ha sido planteada con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento frente a no aplicarlo sobre un escenario donde se usa la técnica de Windowing como preprocesamiento extra. En este caso, además se comparte la misma estrategia de entrenamiento por ambos modelos entrenándolos con un conjunto de training desequilibrado siendo las mismas observaciones en cada una de las épocas con el matiz de que en el caso de la prueba 4 estas observaciones serán recibidas por la red neuronal con ligeras transformaciones (DA) que marcan la diferencia frente a la prueba 5.

En ambas pruebas, se han entrenado los modelos con aproximadamente 87.000 observaciones con una distribución desequilibrada de representación de las clases. En la prueba 4, se aplica DA con el fin añadir diversidad a los datos de entrenamiento y ver que resultado tiene respecto a no utilizarlo como se ha realizado en la prueba 5.

El entrenamiento de ambos modelos fue relativamente similar. En ambas pruebas, aunque no se detectó síntomas de overfitting (como en las pruebas 1 y 2), se estuvieron entrenando hasta que ver que la red ofrecía un progreso de una manera muy lenta. En la prueba 4, el entrenamiento llegó a las 53 épocas lo cual supuso 80 horas de entrenamiento. Por otro lado, en la prueba 5, el entrenamiento llegó a las 95 épocas lo cual supuso 120 horas de entrenamiento.

En comparación con los resultados obtenidos en las pruebas 1 y 2, los resultados han sido bastante mejores con síntomas de avance, sin embargo, este avance se ha producido de una manera muy lenta.

- En la prueba 4:
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos fue:
        - Subconjunto de Clases Negativas: llegó a tener un accuracy de 96% en su última epoch.
        - Subconjunto de Clases EPI: el accuracy llegó a 71% pero el F1 Score 0%.
        - Subconjunto de Clases INP: el accuracy llegó a 77% pero el F1 Score 27%.
        - Subconjunto de Clases INV: el accuracy llegó a 79% pero el F1 Score 30%.

        - Subconjunto de Clases SAH: el accuracy llegó a 76% pero el F1 Score 25%.
        - Subconjunto de Clases SDA: el accuracy llegó a 93% y el F1 Score 85%.
        - Subconjunto Total: se mantuvo en un 81% de accuracy. F1 Score al 30%.

- En la prueba 5, los resultados que se exponen a continuación son de la última época, la época 93. En un principio, se podría llegar a pensar que no es justo comparar un modelo respecto a otro en fases tan dispares de entrenamiento, sin embargo, como se acaba de comentar, el progreso de fue muy lento por lo que los resultados de esta época eran sensiblemente mejores que los de la 53 y en vista a la diferencia de los resultados de los modelos es irrelevante detallar ambas al mismo periodo de ejecución:
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos fue:
        - Subconjunto de Clases Negativas: llegó a tener un accuracy de 95% en su última epoch.
        - Subconjunto de Clases EPI: el accuracy llegó a 70% pero el F1 Score 0.01%.
        - Subconjunto de Clases INP: el accuracy llegó a 95% pero el F1 Score 92%.
        - Subconjunto de Clases INV: el accuracy llegó a 93% pero el F1 Score 88%.
        - Subconjunto de Clases SAH: el accuracy llegó a 92% pero el F1 Score 86%.
        - Subconjunto de Clases SDA: el accuracy llegó a 97% y el F1 Score 96%.
        - Subconjunto Total: se mantuvo en un 91% de accuracy. F1 Score al 66%.

5.3 Pruebas 6-2B, 7-1B y 8

Este último conjunto de pruebas ha sido planteado con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento, cuando se aplica DA sólo a la clase con menor representación y cuando no se aplica DA, todo ello sobre un escenario donde se aplica la técnica de Windowing como procesamiento extra y, además, comparten la misma estrategia de entrenamiento a la hora de distribuir las clases de una manera dinámica y equilibrada en cada una de las épocas del training.

En definitiva, haciendo hincapié sólo en estas tres pruebas se puede ver que las tres comparten todo menos la manera en la que se aplica DA. Por un lado, es interesante ver cómo se comporta DA sobre este escenario, pero lo más interesante es la comparación de los resultados de estas pruebas con los de las pruebas citadas anteriormente con el fin de confirmar si existe mejora o no en cada una de estas condiciones.

La estrategia de distribución de las clases para el entrenamiento es similar a la de las pruebas 1 y 2 y únicamente difieren en el uso de técnica de preprocesado Windowing. Siendo más concretos, la prueba 6-2B es análoga la prueba 2 con única diferencia del uso de Windowing en la presente y, en la prueba 7-1B, sucede lo mismo respecto a la prueba 1.

Por otro lado, las técnicas de preprocesado usadas en las pruebas 4 y 5 son similares, pero difieren en la estrategia de distribución de clases para entrenar.

El entrenamiento de los tres modelos fue relativamente similar ya que los resultados en todas ellas han sido totalmente satisfactorios. En todas ellas, aunque no se detectó síntomas de overfitting, se estuvieron entrenando hasta que ver que la red ofrecía un progreso de una manera demasiado lenta. En la prueba 6-2B, el entrenamiento llegó a las 700 épocas lo cual supuso 84 horas de entrenamiento. Por otro lado, en las pruebas 7-1B y 8, el entrenamiento llegó a las 900 épocas lo cual supuso 90 horas de entrenamiento por cada modelo.

En comparación con los resultados obtenidos en el resto de las pruebas anteriores, estos han sido definitivamente los mejores de todo el estudio. A continuación, se comparan los mejores resultados obtenidos en cada modelo:

- En la prueba 6-2B:
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos para la época 502 (la que mejor resultado proporcionó de las 700) fue:

        - Subconjunto de Clases Negativas: llegó a tener un accuracy de 91%.
        - Subconjunto de Clases EPI: el accuracy llegó a 97% pero el F1 Score 96%.
        - Subconjunto de Clases INP: el accuracy llegó a 94% pero el F1 Score 87%.
        - Subconjunto de Clases INV: el accuracy llegó a 96% pero el F1 Score 93%.
        - Subconjunto de Clases SAH: el accuracy llegó a 94% pero el F1 Score 87%.
        - Subconjunto de Clases SDA: el accuracy llegó a 94% y el F1 Score 89%.
        - Subconjunto Total: se mantuvo en un 95% de accuracy. F1 Score al 84%.

- En la prueba 7-1B:
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos para la época 806 (la que mejor resultado proporcionó de las 900) fue:
        - Subconjunto de Clases Negativas: llegó a tener un accuracy de 94%.
        - Subconjunto de Clases EPI: el accuracy llegó a 98% pero el F1 Score 98%.
        - Subconjunto de Clases INP: el accuracy llegó a 96% pero el F1 Score 94%.
        - Subconjunto de Clases INV: el accuracy llegó a 96% pero el F1 Score 94%.
        - Subconjunto de Clases SAH: el accuracy llegó a 97% pero el F1 Score 93%.
        - Subconjunto de Clases SDA: el accuracy llegó a 96% y el F1 Score 95%.
        - Subconjunto Total: se mantuvo en un 95% de accuracy. F1 Score al 90%.

- En la prueba 8:
    - En cuanto a los resultados de los subconjuntos de Test, la predicción para cada uno de ellos para la época 803 (la que mejor resultado proporcionó de las 900) fue:
        - Subconjunto de Clases Negativas: llegó a tener un accuracy de 93%.

- Subconjunto de Clases EPI: el accuracy llegó a 98% pero el F1 Score 98%.
- Subconjunto de Clases INP: el accuracy llegó a 97% pero el F1 Score 95%.
- Subconjunto de Clases INV: el accuracy llegó a 96% pero el F1 Score 94%.
- Subconjunto de Clases SAH: el accuracy llegó a 96% pero el F1 Score 94%.
- Subconjunto de Clases SDA: el accuracy llegó a 97% y el F1 Score 96%.
- Subconjunto Total: se mantuvo en un 96% de accuracy. F1 Score al 91%.

Capítulo 6. Discusión y Análisis de Resultados

En este capítulo se aborda la discusión sobre el significado de los resultados presentados en el capítulo 5 destacando las ventajas e inconvenientes de cada una de las soluciones evaluadas.

En relación con las pruebas 1 y 2, que trataron de comparar el resultado de aplicar DA de diferente manera bajo un escenario donde no se aplicaba el preprocesado de ventanas (Windowing) y la distribución del conjunto de entrenamiento se hizo de manera dinámica y equilibrada, como se pudo apreciar, no hubo apenas diferencia en sus resultados. Cabe destacar la métrica del F1 Score que fue sensiblemente más baja en el modelo donde se aplica DA a todo el conjunto de datos de entrenamiento (prueba 2) respecto al que sólo aplicaba DA a la clase con menor representación (prueba 1). Esta diferencia fue sólo de un 0.02%.

Llama la atención esta pequeña diferencia de resultados, donde los primeros indicios parece que apuntan a que el uso de DA es sensiblemente contraproducente en este escenario.

Al termino de estas dos pruebas, se concluyó que ambos resultados fueron inesperadamente muy bajos por lo que obligó a seguir continuando en la búsqueda de otras combinaciones con el fin de obtener más pruebas con las que completar esta comparativa esperando encontrar mejores resultados.

Por lo tanto, la siguiente tanda de pruebas, las pruebas 4 y 5, trataron de comprobar la efectividad de DA, pero en un escenario donde se aplicaba el preprocesado de ventanas a las TC y otras estrategias de distribución de las observaciones para el entrenamiento diferentes de las que se aplicaron en las dos primeras pruebas.

En estas dos pruebas se puede apreciar una mejora notable de rendimiento respecto a las pruebas iniciales. El uso de la técnica de preprocesado de ventanas junto con la distribución de los datos de entrenamiento de manera fija y equilibrada han conseguido aumentar las métricas de casi todas las clases respecto a las que se obtuvieron en las pruebas 1 y 2.

Tal y como se comentó en capítulos anteriores, existe un claro ejemplo donde se demuestra la importancia de no fiarse únicamente de la métrica de la exactitud o accuracy. En la prueba 4, más en concreto, en los resultados del subconjunto de test para la clasificación Epidural, la

exactitud ronda el 75% y, sin embargo, el F1 Score es del 0%. Esto significa que el modelo es incapaz de clasificar adecuadamente las hemorragias etiquetadas como Epidural pero sí que ha sido capaz de identificar correctamente que los otros subtipos de ICH no estaban presentes realmente, de ahí el porqué del 75% de exactitud. En este caso, para los 300 casos etiquetados con ICH Epidural de este subconjunto, el modelo no ha clasificado a ninguno como ICH de subtipo Epidural, lo cual está muy alejado del objetivo de este trabajo. En cuanto al resto de subtipos han sido clasificados con un F1 Score entre un 71% el más bajo y un 93% el que más representación tiene, obteniendo un 30% en el Total.

Por otro lado, en el caso de la prueba 5, a pesar de que ésta tiene mejores resultados en comparación con la prueba 4, también se puede ver el mismo problema a la hora de clasificar la ICH Epidural (F1 Score 0.01%). En cuanto al resto de resultados se aprecia una mejora muy notable en exactitud y F1 Score de la mayoría de las clases positivas, superando el 90% y rondando el 89% respectivamente, obteniendo un 66% en el Total.

Por lo tanto, es en la clase Epidural donde ambos modelos suspenden. Esto es debido a la poca representación que existe entre la clase epidural y las restantes. En cada epoch de 87.000 observaciones, cuando de la clase epidural se han procesado 1.047 observaciones, de la clase Subdural se han visto más de 20.000. Esto hace que los pesos de la red se vean más afectados de cara a detectar ICH de clases con mayor representación por lo que la red no llega a ser buena generalizando las ICH de tipo Epidural y bastante buena a la hora de clasificar ICH de tipo Subdural.

Como se puede apreciar en los resultados de estas dos pruebas, la diferencia de estos ya empezó a ser notable. En este nuevo escenario, se pudo ver que el uso de DA (prueba 4) parecía que tenía un efecto bastante negativo respecto a si no se usaba (prueba 5), una diferencia del F1 Score Total de un 36% entre ambas, la cual era muchísimo mayor del que se tuvo con las 2 primeras pruebas (0.02%). Sorprendentemente, incluso utilizando técnicas de DA, que se podría llegar a pensar que (a priori) aportan mejores resultados utilizando DA que si no, se vuelve a comprobar que el resultado es peor usando DA para todo el conjunto que usándolo para sólo una clase o no usándolo directamente.

Todo esto podría verse explicado debido a que DA, sobre todo, funciona muy bien cuando el conjunto de datos es pequeño y en este caso, 87.000 observaciones no lo es tanto. Por otro lado, también se puede llegar a pensar que las transformaciones que DA ha realizado son tan exageradas que podría ser que el modelo las interpretase como observaciones totalmente dispares y le cueste llegar a aprender todos los mismos, e incluso, necesitando un tiempo mayor de entrenamiento del que se ha dedicado.

Por último, tratando de comparar el resultado de la prueba 4 con los resultados de la prueba 2, donde ambas usaban técnicas de DA aplicadas a todas las clases, se puede apreciar que hay dos diferencias notables: la estrategia de la distribución de las clases en el conjunto de entrenamiento y el uso o no de Windowing. Debido a heterogeneidad de ambas técnicas, ha sido necesario seguir investigando con el fin de encontrar cual es la pieza clave dentro de este rompecabezas para encontrar la mejor predicción en la detección de ICH y la correcta clasificación de sus subtipos. Llama la atención ver que a pesar de la cantidad de tiempo de entrenamiento y del gran volumen de los datos aportado gracias a DA, el modelo haya conseguido, sobre todo en las pruebas 1 y 2, unos resultados tan bajos.

Como conclusión y analizando los resultados vistos hasta ahora, da la sensación de que la técnica de DA no consigue aportar tanto cómo se esperaba. Por otro lado, la estrategia de distribución de las clases de un conjunto de entrenamiento fijo y sobre todo desbalanceado parece que ha aportado mejores resultados, pero, como se puede apreciar, aquellas clases con menor representación tienen una penalización muy elevada que obliga a seguir investigando. Probablemente, el motivo de estos resultados vistos hasta ahora radiquen en la incapacidad que tiene la red de poder identificar los detalles de la imagen relacionados con la ICH y la anatomía del cerebro debido a la carencia de técnicas de preprocesado de ventanas que permitirían identificarlos.

Por ello, con la siguiente tanda de pruebas, pruebas 6-2B, 7-1B y 8, trataron de sacar de dudas esta última hipótesis, repitiendo, por un lado, la prueba 1 y 2 pero usando la técnica de preprocesado de ventanas. Además, al ver el pésimo efecto de DA hasta ese momento, se decidió añadir una tercera prueba, similar a la 6-2B y 7-1B pero sin el uso de esta técnica.

Satisfactoriamente, el resultado de estos tres modelos fue un éxito respecto a los resultados obtenidos en pruebas anteriores. Además, la diferencia de resultados entre las tres pruebas ya no es tan evidente. Únicamente hay una diferencia moderada entre la prueba 6-2B, que es la que aplica DA a todas las clases del modelo, y las otras dos que, o no aplica DA, o sólo la aplican en la clase con menor representación. Con esto, se vuelve a reafirmar que para este tipo de problemas DA no aporta gran valor, incluso, todo lo contrario, es contraproducente usarlo.

En cuanto a las pruebas 7-1B (DA en Epidural) y 8 (Sin DA), la diferencia es prácticamente nula, siendo igual o sensiblemente mejor, con un incremento de 0.01% de F1 Score en prácticamente todas clases, el modelo que no presenta uso de DA.

De esta manera, se ha comprobado la importancia del uso de las técnicas de preprocesado de ventanas junto con el redimensionado a 256x256 y la normalización, además del cuidado

que hay que tener a la hora de elegir una buena estrategia para servir los datos a la red de una manera equilibrada y dinámica y evitando en todos los casos el uso de DA. Con todo esto, se ha conseguido demostrar el efecto que tiene el usar o no este tipo de técnicas.

Por último, como se comentó en capítulos anteriores, se compara los resultados obtenidos del mejor modelo, prueba 8, frente los resultados recogidos del modelo de referencia trabajo de H. Ye (Ye, 2019) sobre el que se decidió comparar los resultados. Se puede apreciar que en todos los subtipos de ICH se consigue mejorar notablemente las métricas de la exactitud y del F1 Score:

Tabla 4
Comparativa de resultados

	Ye 2019		**Prueba 8**		**Diferencia**	
	Accuracy	F1 Score	**Accuracy**	**F1 Score**	Accuracy	F1 Score
EPI	0,96	0,72	**0,98**	**0,98**	0,02	0,26
INP	0,9	0,93	**0,97**	**0,95**	0,07	0,02
INV	0,91	0,87	**0,96**	**0,94**	0,05	0,07
SAH	0,83	0,78	**0,96**	**0,94**	0,13	0,16
SDA	0,94	0,84	**0,97**	**0,96**	0,03	0,12

Tabla 4 Comparativa de resultados entre el modelo presentado en el trabajo de H. Ye y el mejor modelo presentado en este trabajo.

Capítulo 7. Conclusiones y Trabajo Futuro

A continuación, se reúnen las conclusiones derivadas de este Trabajo de Fin de Master en base a los objetivos específicos previamente mencionados en el capítulo 3.

Los primeros objetivos consistían en la revisión sistemática de la literatura actual, esta tarea se abordó de forma minuciosa en el capítulo 2, concluyendo de que había una falta de claridad en la efectividad del uso de algunas técnicas de preprocesado de datos y métodos de balanceo de estos por lo que se planteó una comparativa para ver cómo de efectivos eran este tipo de técnicas en el ámbito de la detección de ICH y sus subtipos. En base a ciertos trabajos de relevancia que se tuvieron en consideración, se decidió poner en evaluación la efectividad las técnicas de preprocesado de Windowing y diferentes métodos de balanceo de datos como el uso Data Augmentation para la resolución del citado problema.

De esta manera, tal y como se indicó en el objetivo número 3, se definió un banco de pruebas donde se trataba de comprobar la efectividad de estas. Siguiendo los pasos del objetivo 4, tras el entrenamiento de los modelos y la posterior evaluación de cada uno con los mismos conjuntos de tests, se obtuvieron los resultados de cada una de las pruebas.

En base al último objetivo y teniendo en cuenta estos resultados, se realizó una comparación de estos y se pudo demostrar que el uso de la técnica de preprocesado de ventanas junto con la manera en que se alimenta la red a la hora de entrenar el modelo con unos datos distribuidos de una manera equilibrada y dinámica en cuanto al grado de representación de las clases, son la pieza fundamental en la capacidad de generalización y una elevada predicción del modelo. Además, también se demostró que el uso de técnicas de balanceo de datos generando datos sintéticos (Data Augmentation) como entrada al modelo, tenía un efecto contraproducente por lo que se ha desaconsejado el uso del mismo.

Por lo tanto, las opciones que se acaban de comentar son las técnicas y métodos recomendados a la hora de implementar modelos de diagnóstico de ICH y la clasificación de sus subtipos. No obstante, debido a la naturaleza de los datos y la versatilidad de estas técnicas y a la sencillez de su utilización, también se puede utilizar para la detección de otro tipo de enfermedades que pudiera detectar un radiólogo por medio de Tomografías Computarizas, tales como cancer, alzheimer, ulceras, lesiones de cualquier tipo, otro tipo de hemorragias, etc. … ya que no difieren notablemente en los procedimientos que se tienen que seguir para su diagnóstico.

Todo lo comentado en esta conclusión, así como, durante toda la fase comparativa experimental, responde a la pregunta planteada durante todo el trabajo de ¿qué técnicas, de las opciones disponibles, logra los mejores resultados?

En cuanto a las contribuciones aportadas por este trabajo se pueden resumir de la siguiente manera:

- Se ha analizado y resumido de una manera detallada todo el estado del arte del diagnóstico de las hemorragias intracraneales y la identificación de sus subtipos a través de tomografías computarizadas a nivel clínico y el diagnóstico automático a través de algoritmos de inteligencia artificial.
- Se ha proporcionado las recomendaciones necesarias para aplicar a los modelos de detección automática en relación con las técnicas de preprocesado y métodos de balanceo de datos que se deben utilizar o no.
- Se ha demostrado a través de uno de los modelos evaluados que es posible mejorar los resultados de predicción de trabajos recientes en relación con el diagnóstico automático de las ICH y la clasificación de los subtipos.

Es importante concluir con que no se puede afirmar rotundamente de que este tipo de modelos se puedan llevar a producción y que actúen de manera autónoma fácilmente. A pesar de haber obtenido unos resultados bastante destacables, incluso en comparación con los últimos trabajos publicados sobre este campo, los modelos aún tienen un largo camino de mejora hasta llegar a un 100% de exactitud.

No obstante, el rendimiento de este tipo de modelos podría comprobarse de otra manera. Partiendo de la idea de que el uso de Inteligencia Artificial en medicina busca reducir el error humano, se podría valorar cómo de bueno es un modelo a partir de una comparación directa con radiólogos. En este escenario, un modelo podría no necesitar resultados prácticamente perfectos, sino simplemente igualar o superar a los radiólogos. Además, su uso se plantearía como un complemento al trabajo de los radiólogos, y no como un sustituto, permitiendo al profesional tener siempre la última palabra, pero permitiéndole reducir sus errores de forma notable. Así, la implementación de estos modelos por parte de profesionales sin conocimiento profundo de la enfermedad podría ser más razonable, aunque siempre sería necesario un procedimiento de evaluación basado en la comparación directa con radiólogos.

Por esto y debido a la escasez de tiempo que se ha permitido invertir en este Trabajo de Fin de Master, se plantean las diferentes líneas de futuro que podrían ayudar a mejorar los resultados de modelos de este tipo.

En cuanto al balanceo de datos, se podría usar técnicas de estratificación de los datos con el fin de realizar un undersampling de una manera más cuidadosa eliminando observaciones que no fueran tan representativas y evitando hacerlo con aquellas que tuvieran más valor.

En cuanto al preprocesamiento de los datos, sería interesante hacer algún tipo de segmentación de las imágenes con el fin de extraer únicamente el cráneo, de tal manera que se reduciría notablemente el esfuerzo computacional a la hora de entrenar el modelo.

Por otro lado, sería también muy interesante cambiar la última capa softmax del modelo por una capa multi-class que nos permitiera dar una distribución de probabilidad de cada una de las clases con el fin de clasificar correctamente a los pacientes que tuvieran más de un tipo de ICH.

Por último, también se debe tener en cuenta los avances en el mundo del Deep Learning en general con el fin de abrir nuevas vías de investigación para esta línea de investigación. El mundo de la Inteligencia Artificia avanza muy rápido y cada día se publican nuevas técnicas o arquitecturas que mejoran enormemente a las anteriores, y que permiten el entrenamiento de redes más complejas, capaces de conseguir una convergencia mucho más rápida, con mejores resultados, ...

Bibliografía

A. Qureshi, S. T. (2001). Spontaneous intracerebral hemorrhage. *Med Line*, 1450-1460.

American Society of Neuroradiology. (1965). Obtenido de ASNR: https://www.asnr.org/

Amidi, S. (2018). *Standford University*. Obtenido de https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly

Arbabshirani, M. R. (2018). Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine, 1*(1), 1-7.

Buzug, T. M. (2011). *Computed tomography. In Springer Handbook of Medical Technology.* Springer, Berlin, Heidelberg: Springer.

C.R. Becker, M. R. (2008). *Multislice CT.* Springer.

Chollet, F. (2015). *Keras*. Obtenido de GitHub: https://github.com/fchollet/keras

Chollet, F. (2015). *Keras*. Obtenido de github.com/keras-team/keras

D. Escudero Augusto, L. M. (2008). Actualización en hemorragia cerebral espontánea. *Medicina Intensiva, 32*(6), 282-295. doi:10.1016/S0210-5691(08)70956-2

Darias Plasencia, Ó. (2019). *Medicina personalizada: comparativa de técnicas para el diagnóstico automático del Alzheimer.* Logroño (España): Universidad Internacional de la Rioja.

Díaz-Guzmán J, B.-P. F.-L. (2008). Prevalence of Stroke and Transient Ischemic Attack in Three Elderly Populations of Central Spain. *Neuroepidemiology, 30*(4), 247-53.

Ding, Y. S. (2018). A Deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain. *Radiology, 290*(2), 456-464.

Esteva, A. (2017). Dermatologist-level classification of skin cancer. *Nature, 542*(7639), 115.

Fraile, F. (2004). *Imagen radiol´ogica: principios f´ısicos e instrumentaci´on.* Masson.

Franco Martinez, E. A. (2011). Analisis digital de imágenes tomográficas sin contraste para la búsqueda de tumores cerebrales. *Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional*.

Goodfellow, I. B. (2016). *The Deep Learning Book.* Cambridge (Estados Unidos): The MIT Press.

Google, B. (2020). *TensorFlow 2.1*. Obtenido de github.com/tensorflow/tensorflow

Greenspan, H. (2016). *Special Section on Deep Learning in Medical Application*. Obtenido de IEEE: http://www.ieee-tmi.org/deep-learning-in-medical-application.pdf

Gulshan, V. (2016). Development and validation of a deep learning . *JAMA - J. Am. Med. Assoc.*, 2402-2410.

Jupyter, P. (2015). *Jupyter Project*. Obtenido de jupyter.org

K. He, X. Z. (2015). Deep Residual Learning for Image Recognition.

Kidwell, C. S. (2004). Kidwell, Chelsea S., et al. "Comparison of MRI and CT for detection of acute intracerebral hemorrhage. *Jama 292*, 1823-1830.

Klein, A. (2009). Evaluation of 14 nonlinear deformation algorithms. *Neuroimage, 46*(3), 786-802.

Krizhevsky, A. S. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems, 25*(2). doi:10.1145/3065386

L. Wan, M. Z. (2013). Regularization of Neural Networks using DropConnect. *in International Conference on Machine Learning*, 1058–1066.

Lacerda Gallardo, Á. J. (2000). Mortalidad por hemorragias intracerebrales espontáneas: Estudio clinicopatológico. *Revista Cubana de Cirugía, 39*(2), 97-102.

Láinez JM, P. A.-F. (2002). *Guía para el tratamiento y prevención del Ictus*. Obtenido de Sociedad Española de Neurología: http://www.sen.es/profesionales/ictus.htm

LeCun, Y. B. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. doi:10.1109/5.726791

M. Ariesen, S. C. (2003). Risk factors for intracerebral hemorrhage in the general population: a systematic review. *Stroke*, 2060-2065. Obtenido de http://dx.doi.org/10.1161/01.STR.0000080678.09344.8D

Maciej A. Mazurowskia, P. A. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 427-436.

Madrigal Mesa, M. D. (Junio de 2016). *Atención de Enfermería en el Ictus.* (U. d. Jaén, Ed.) Jaen. Obtenido de https://hdl.handle.net/10953.1/2784

Martínez, F. &. (2011). *Análisis digital de imágenes tomográficas sin contraste para la búsqueda de tumores cerebrales.* Unidad Zacatenco: In Departamento de Computación, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional.

Nielsen, M. (12 de 2019). *Neural Networks and Deep Learning*. Obtenido de Neural Networks and Deep Learning: http://neuralnetworksanddeeplearning.com/index.html

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 2018). San Francisco, CA, USA: Determination press.

Nogueira, F. (2016). *Over-sampling*. Obtenido de imbalanced-learn: https://imbalanced-learn.org/stable/over_sampling.html

Prieto, M. H. (2005). *Manual pr´actico de TC. Editorial Médica Panamericana.*

Radiological Society of North America. (1915). Obtenido de RSNA: https://www.rsna.org/

Rawat, W. &. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation, 29*(9), 2352-2449.

Reyes, A. K. (2015). Fine-tuning Deep Convolutional Networks for Plant Recognition. *CLEF* , 467-475.

Rosa, G. P. (2015). Fine-tuning convolutional neural networks using harmony search. *Iberoamerican Congress on Pattern Recognition* , (págs. 683-690). Springer, Cham.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Shin, H. C. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *In International workshop on simulation and synthesis in medical imaging*, 1-11.

Srivastava, N., Hinton, G., Krizhevsky, A. S., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*, 1929-1958.

Sudeep, K. K. (2017). Preprocessing for image classification by convolutional neural networks. *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. doi:10.1109/RTEICT.2016.7808140

Szegedy, C. e. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, págs. 1-9.

T. Brott, J. B. (1997). Early hemorrhage growth in patients with intracerebral hemorrhage. *Stroke, 28*, 1-5.

V. Vanhoucke, C. S. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, págs. 2818-2826.

Woods, R. P. (1993). MRI-PET registration with automated algorithm. *Journal of computer assisted tomography, 17*, 536.

Ye, H. G. (2019). Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European radiology, 29*(11), 6191-6201.

Zhou, Z. S. (2017). Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7340-7351.

Zisserman, K. S. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Prepr. arXiv*, págs. 1409-1556.

Comparativa de Técnicas de Preprocesamiento y Entrenamiento para Detección de Hemorragias Intracraneales

Fernando Arroyo Rubio

Universidad Internacional de la Rioja, Logroño (España)

Fecha 26 febrero 2020

Resumen

En las últimas dos décadas, el mundo de la medicina está teniendo un notable interés por el uso del procesamiento de imágenes con Inteligencia Artificial para la detección de numerosas enfermedades y lesiones que tienen lugar en el ser humano. Específicamente, las técnicas de Machine Learning y Deep Learning han dejado ver sus bondades para ayudar a reducir el error humano y apoyar a que los diagnósticos sean más acertados de los que se obtienen a través de métodos tradicionales tratando de encontrar patrones ocultos dentro de las imágenes a procesar. En este Trabajo de Fin de Máster, se aplican diferentes técnicas de preprocesamiento de imágenes y algoritmos de entrenamiento para la detección e identificación de distintos tipos de Hemorragias Intracraneales (ICH) a través de tomografías computarizadas (TC). Todo esto se implementa en un escenario experimental con el objetivo de conocer qué técnica y algoritmo ofrece un mejor resultado. Los resultados demuestran que el uso de redes neuronales convolucionales, técnicas de preprocesado como Windowing y el empleo de métodos de balanceo de datos permiten elaborar modelos bastante precisos con resultados que superan, incluso, trabajos de referencia de la actualidad. El modelo con mejor resultado logra valores > 0.91% en F1 Score y de >96% en exactitud de media en todos los subtipos. Este modelo es capaz de clasificar correctamente con unos niveles altos de precision pudiendo llegar a ser un sistema de soporte para médicos y radiólogos en hospitales.

Palabras Clave

Análisis Imagen Médica, Aprendizaje Profundo, Hemorragia Intracraneal, Redes Neuronales Convolucionales, Preprocesamiento Imagen.

I. Introducción

Las Hemorragias Intracraneales (ICH), son un problema de salud grave originados por una lesión vascular provocada en el interior del cerebro. Este tipo de dolencia afecta cada año a más de 130.000 españoles, de las cuales la mitad sufren alguna limitación en sus funciones a causa de las ICH. Es la segunda causa de muerte en nuestro país, la primera entre las mujeres [1].

Cuando un sujeto presenta síntomas de un accidente cerebrovascular, es de vital importancia la rápida intervención del personal médico para realizar un examen médico e intervenir lo más rápidamente posible. De lo contrario, graves secuelas podrían derivarse al enfermo e incluso el resultado podría ser fatal llegándose a producir la muerte.

En contexto, la Inteligencia Artificial, específicamente con el desarrollo del Deep Learning, han permitido el desarrollo de sistemas para la detección de ICH a partir de imágenes médicas. Los estudios indican que se han logrado resultados prometedores [2][3] en cuanto a que pueden apoyar el diagnóstico de los especialistas, aportando más evidencias. La importancia de la disponibilidad en los centros de salud de una segunda opinión confiable, "virtual" y de bajo coste, que es entrenada por neurorradiólogos y que puede ayudar a que los servicios médicos expertos y no expertos sean más eficientes y confiados.

Además, es muy importante conseguir un sistema altamente eficiente y sobre todo fiable ya que contribuiría a descargar notablemente la carga de trabajo de los hospitales. Un sistema artificial que podría actuar como un primer filtro de las tomografías computarizadas a analizar y en caso de que éstas presenten hemorragias, alertar al equipo de atención altamente cualificado.

Por lo tanto, entendiendo la importancia de un rápido diagnóstico de las ICH junto a las razones que acabamos de explicar, este trabajo se centró en la detección de las ICH con el fin de agilizar su diagnóstico y mejorar los sistemas y procedimientos hoy día existentes en el sector médico. Por ello, es importante preguntarse, ¿cuál de las opciones ofrece mejor rendimiento, no sólo a la hora de desplegar sistemas en entornos reales sino también a la hora de profundizar en la investigación? ¿qué técnicas, de las disponibles, logra los mejores resultados? Estas son unas de las preguntas de investigación que han orientado este TFM y que se intenta resolver tratando de encontrar cuales son las mejores técnicas para conseguir mejorar la capacidad predictiva del modelo.

II. Estado del Arte

A lo largo de estos últimos años se han realizado innumerables esfuerzos usando Aprendizaje Automático. Gracias al Deep Learning, se han conseguido innumerables avances, en las que se proponen diversas técnicas para tratar de hacer frente a la detección de ICH a través de imágenes médicas, principalmente o, al menos las que más éxito han tenido, bajo el paraguas de las redes neuronales profundas. [4][2][3].

Arbabshirani, en su trabajo [3], demostró un impacto positivo en la optimización del flujo de trabajo en radiología construyendo un

algoritmo de inteligencia artificial que ayudaba a priorizar las listas de trabajo de dicha área en base a la criticidad del paciente con el fin de reducir el tiempo de atención en un 96%. Además, conseguía identificar ICH muy sutiles pasadas por alto por algunos radiólogos. Para ello, empleó una red neuronal convolucional profunda entrenándola con más de 2 millones de tomografías computarizadas de diferentes tamaños debido a las diferencias de tamaño de los cráneos, por lo que se tuvo que hacer un redimensionado o resizing a 256x256. Además, se emplearon técnicas de Windowing aplicando un filtro de cerebro para identificar las hemorragias (centrada en 40HU y ancho de 80HU). Por último, para aumentar los datos de entrenamiento y conseguir un conjunto de datos más equilibrado, se aplicó DA sobre todo el conjunto de entrenamiento donde se realizaron aproximadamente entre 20 y 80 observaciones aumentadas de cada estudio negativo y estudio positivo, respectivamente. Se aumentó aplicando una traslación horizontal y vertical aleatoria (±0-20 píxeles), rotación (±0-15°) y volteado (horizontal).

Por otro lado, el trabajo de Darías[4], publicado recientemente, a pesar de no ser un estudio enfocado en la dirección del diagnóstico de ICH directamente, ha trabajado en una línea muy similar tratando de detectar Alzheimer a través de imágenes médicas de RMI a través del uso de técnicas de Deep Learning. Este trabajo se basó principalmente hacer una comparativa del resultado proporcionado por dos de las redes neuronales profundas más utilizadas hoy día InceptionV3 y ResNet50 para identificar la enfermedad citada. Darías, debido a que el data set utilizado no era muy grande, se encargó de utilizar técnicas de fine-tuning sobre ambas redes con el fin de aprovechar el preentrenamiento de ambas y conseguir mejor resultado. Como técnicas de preprocesado utilizó técnicas de redimensionamiento y normalización de imágenes con el fin conseguir un entrenamiento óptimo de la red. Este estudio ha sido muy importante en el desarrollo de esta investigación ya que plantea, en su apartado de líneas futuras, una serie de recomendaciones sobre qué caminos se deben seguir a fin de seguir avanzando en la investigación proponiendo hacer uso de técnicas como Windowing o Data Augmentation.

Por último, el trabajo de H. Ye [2] publicado recientemente hace menos de un año, es uno de los trabajos de más relevancia en el mundo de la detección hemorragias intracraneales por lo que también se ha tenido muy en consideración. En este trabajo, se evalúa la capacidad de predicción de ICH sobre tomografías computarizadas y sus subtipos usando una combinación de red neuronal convolucional (CNN) con una red neuronal recurrente (RNN) demostrando que incluso la capacidad de precisión media era incluso mejor que la de algunos radiólogos junior. Empleó técnicas de preprocesamiento para poder entrenar al modelo. Por un lado, utilizó resizing con el objetivo de conseguir una reducción del tamaño de las TC de 512x512 a 256x256 reduciendo el uso de memoria de la GPU. Por otro lado, empleó Windowing con 3 ventanas diferentes (huesos, tejidos, blandos, vasos sanguíneos, …) Además, utilizó Data Augmentation: con el fin de reducir overfitting y equilibrar el desbalanceo de datos aplicando técnicas de rotación (0-180 grados), reescalado (90%-110%), deslizamiento (10%), recortes (10%) y volteado vertical y horizontal.

Tras el análisis de la literatura actual en base a los conceptos clínicos técnicos y del estado del arte expuestos en todo este apartado, se puede concluir que existen técnicas cuyo aporte es claro e indiscutible, pero, sin embargo, existen otras que no queda del todo demostrada su efectividad por lo que no se puede concluir completamente qué técnicas de preprocesamiento o métodos de entrenamiento son los que mejor resultado acaba proporcionando.

En base a lo explicado en este apartado, el redimensionamiento o resizing puede catalogarse como una técnica cuyo aporte es indiscutible, se identifica que el uso generalizado de esta técnica en todas las literaturas analizadas, [4][2][3] estableciendo las imágenes disponibles con unas dimensiones fijas y comunes. Además, otra de las técnicas donde el aporte también es evidente es la técnica de la normalización o whitening con el fin de entrenar los modelos de una manera más eficiente y rápida para conseguir una mejor convergencia de la red. En cuanto a los modelos utilizados, también ha tenido un uso generalizado las redes neuronales, más en concreto, el uso de redes neuronales previamente entrenadas con ImageNet y hacer el uso de técnicas como Transfer Learning y fine-tuning, [4][2][5][6]. Por lo tanto, este trabajo se centra en el uso de esta técnica.

Por otro lado, en cuanto a los métodos de balanceo de clases no parece que sea muy generalizado el uso de técnicas a priori que permitan equilibrar el conjunto de datos para el entrenamiento o, al menos, no se menciona sí se ha realizado o no. La única que técnica más extendida es DA pero es complicado determinar el grado de aportación que aplica esta técnica a los resultados de los modelos. Por lo tanto, este punto es objeto de estudio en este trabajo.

En relación con otra de las técnicas que, en base a la literatura analizada, parece tener una gran importancia y un uso extendido es la técnica de Ventanas de TC o Windowing. En los trabajos de Ye y Arbasbshirani [2][3] es utilizada esta técnica aplicando diferentes ventanas con el fin de extraer las características más importantes de las TC.

Finalmente, en base a las recomendaciones de líneas futuras que hizo Darías [4], se ha establecido también como objeto de estudio el uso de técnicas de preprocesamiento como Windowing y métodos de balanceo de datos como Data Augmentation con el fin de comprobar cómo de efectivo es el uso de las mismas. Además, se establecerá como punto de referencia los resultados ofrecidos por el trabajo de Ye [2] que, en base a la literatura inspeccionada, ha sido el estudio que mejor resultado ha proporcionado hasta la fecha.

III. Objetivos y Metodología

El objetivo general de este estudio es hacer recomendaciones sobre el uso de diferentes técnicas que ayuden al entrenamiento de una red neuronal con el fin de mejorar la detección de hemorragias intracraneales sobre tomografías computarizadas realizando una comparación experimental de las técnicas expuestas. Esto implica

1. Revisión sistemática de la literatura actual con relación a la detección de hemorragias intracraneales a nivel clínico.

2.Identificar los modelos, técnicas de preprocesamiento y métodos de balanceo de datos para el entrenamiento de una red neuronal que mejor resultado han proporcionado para la detección automática de hemorragias intracraneales.

3. En base a las carencias identificadas en los objetivos anteriores, definir un banco de pruebas e implementar los modelos considerados que pueden ayudar a mejorar la capacidad de predicción de un sistema artificial de detección de ICH.

4. Entrenar los modelos, analizando sus resultados y la capacidad de predicción de cada modelo diferentes tipos de hemorragia.

5. Comparar los resultados obtenidos por los diferentes modelos entrenados y emitir las recomendaciones oportunas.

Para ello, la metodología empleada se dividió en dos fases principales: la fase exploratoria y la fase comparativa experimental.

La fase exploratoria se llevó a cabo para identificar aquellas técnicas que más se usan en el diagnóstico automatizado de las ICH. En particular se realizaron las siguientes actividades en esta fase:

1. Elaboración de una revisión sistemática de la literatura actual que permitió una mejor comprensión de las ICH y su diagnóstico a nivel clínico.

2. Elaboración de una revisión sistemática de literatura que permitió una mejor comprensión de la aplicación del Aprendizaje Automático en el diagnóstico de las ICH actualmente.

3. Identificar los modelos, técnicas de preprocesamiento y métodos de balanceo de datos para el entrenamiento de una red neuronal que mejor resultado han proporcionado para la detección de ICH a través de una revisión de la literatura actual. Como resultado, se proporciona información sobre los modelos que serían más indicados a la hora de desplegar un sistema de diagnóstico de ICH en un entorno real, y sobre los principales métodos de preprocesamiento y técnicas balanceo de datos necesarios para ponerlos en práctica.

Por otro lado, la fase descriptiva se realizó con el objetivo de describir el funcionamiento y comportamiento de los modelos seleccionado en la fase anterior y comparar de manera experimental sus resultados. Esto implica:

1. Definición de un banco de pruebas con los diferentes modelos, técnicas y métodos identificados en las conclusiones del estado del arte con el fin de preparar la comparativa.

2. Implementación de los modelos usando las diferentes técnicas y métodos definidos en el banco de pruebas. Ajuste o balanceo de los conjuntos de datos y entrenamiento de los modelos en base a las directrices indicadas en la definición de cada una de las pruebas.

3. Obtención de resultados del entrenamiento y la evaluación de cada uno de los modelos.

4. Valoración final teniendo en cuenta los resultados obtenidos, analizándolos y comparando los mismos y contrastando con la información extraída durante la fase exploratoria.

5. Realizar recomendaciones pensadas para profesionales de Inteligencia Artificial sin conocimientos profundos de medicina. Partiendo de las conclusiones del punto anterior, se busca proporcionar una base sobre la que partir a la hora de construir nuevos modelos de diagnóstico o la expansión de este.

IV. Contribución

El objetivo de esta comparativa es averiguar qué técnicas y qué métodos de balanceo de datos, extraídos de la fase exploratoria, proporcionan mejor un resultado para la detección de ICH y la clasificación de subtipos en TC cerebrales. Para ello, se han definido una serie de pruebas que utilizan estas técnicas con el fin de tratar de comparar los resultados entre sí y comprobar cuál es la que mejor resultado proporciona. Algunas de las técnicas, sobre todo aquellas que su aportación era muy evidente, han sido utilizadas en todas las pruebas. Por el contrario, el resto de las técnicas cuyo aporte necesitaba ser evaluado, se utilizan en algunas pruebas y en otras se ha descartado con el fin de comprobar su aporte y efectividad a la hora de detectar ICH y clasificar los subtipos.

Conjunto de datos

La pieza más importante en un problema de clasificación con IA son los datos. Para esto se ha utilizado un conjunto de datos proporcionado por la Sociedad Radiológica de América del Norte o, en inglés, Radiological Society of North America (RSNA) (Radiological Society of North America, 1915). De los datos cedidos, se han obtenido 674.257 tomografías computarizadas de más de 25.000 exámenes clínicos realizados a más de 17.079 pacientes. Los datos fueron proporcionados totalmente anonimizados y cada tomografía tenía asignada 6 etiquetas dónde se establecía valor ‘1’ o ‘0’ en caso de ser padecer o no la hemorragia asociada: ANY, paciente con algún tipo de ICH; EPI, paciente con un tipo de ICH Epidural; INP, paciente con un tipo de ICH Intraparenquimatosa; INV, paciente con un tipo de ICH Intraventricular; SAH, paciente con un tipo de ICH Subaracnoidea; SDA, paciente con un tipo de ICH Subdural. El conjunto de datos está desbalanceado considerablemente. A pesar de las recomendaciones de Mazurowskia [7], con el fin de que el reparto entre las clases positivas frente a las negativas sea más equilibrado, se ha decidido aplicar la técnica de undersampling a las tomografías de la clase negativa pasando de 577.154 imágenes o cortes a 57.983. De este modo, el reparto entre clases positivas y negativa se quedó en un 54% frente a 46%. Esta decisión ha sido tomada debido a que es preferible que el modelo se equivoque en el caso de que una imagen etiquetada como negativa se prediga como positiva en vez de lo contrario. Sería más preocupante si un paciente con ICH se le clasificase como un paciente sano.

Finalmente, la distribución final que se ha quedado en el conjunto de datos se presenta a continuación en la Tabla I. Además, se aprovecha para indicar la distribución entre conjunto de entrenamiento, validación y test:

Tabla I

Distribución del conjunto de datos

Class	Train	Val	Test	Total
NEG	40.401	5.774	11.808	**57.983**
EPI	1.047	149	301	**1.497**
INP	9.752	1.380	2.800	**13.932**
INV	6.239	883	1.792	**8.914**
SAH	10.301	1.458	2.958	**14.717**
SDA	20.236	2.863	5.810	**28.909**
Total	**87.976**	**12.507**	**25.469**	**125.952**

Distribución del conjunto de datos proporcionado por la RSNA (Radiological Society of North America, 1915) a utilizar en este estudio

Pre-Entrenamiento

Antes de entrenar el modelo, se requiere un cierto preprocesamiento antes de que los datos estén listos para ser consumidos por la red neuronal. La primera técnica aplicada es el Resizing a 256x256 a todas las imágenes médicas que se utilizan en todos los modelos, tal y como dictaba el trabajo de Ye [2], con el fin de tratar de agilizar el entrenamiento ya que permite reducir bastante el uso de memoria de la GPU. Tal y como se concluyó en el apartado II, en base a los trabajos estudiados [4] [2] [3] quedó demostrado la aportación del redimensionamiento es indiscutible por lo que se ha mantenido este método en todas las pruebas que realizadas en este estudio.

La siguiente técnica, Windowing, se comprueba cuanto de efectiva es esta técnica. En base a los trabajos estudiados [2] [3] el objetivo de esta técnica es ser capaces de identificar diferentes tipos de tejidos dentro de la TC.

Las tomografías del dataset han sido proporcionadas por la RSNA en el formato DICOM. DICOM es el estándar que más se utiliza para almacenaje e intercambio de imágenes médicas obtenidas por TC. Este formato, además de contener la imagen médica en sí, contiene muchos metadatos tales como: el tamaño de los píxeles, la longitud de un píxel en cada dimensión del mundo real, …

La unidad de medida en las tomografías es la Unidad Hounsfield (HU), que es una medida de la radiodensidad. Los escáneres TC están cuidadosamente calibrados para medir esto con precisión. Desafortunadamente, los valores devueltos no están en esta unidad. Por lo tanto, es necesario hacer unos calculos para poder trabajar en

esta unidad.

Algunos escáneres tienen límites de escaneo cilíndricos, pero la imagen de salida es cuadrada. Los píxeles que caen fuera de estos límites obtienen el valor fijo de -1024, por lo que el primer paso será fijar estos valores a -1024, valor que corresponde al aire como se ha explicado en el apartado II (Franco Martinez, 2011).

Seguidamente, se multiplica el valor de cada píxel por la pendiente de reajuste y se suma con la intercepción. Ambos valores se pueden extraer de los metadatos del fichero DICOM a través de los campos RescaleSlope y RescaleIntercept respectivamente.

Una vez se ha convertido la imagen a HU, se procede a aplicar las respectivas ventanas con el fin de extraer únicamente los tejidos más relevantes para la identificación de ICH ya que cada tipo de tejido está asociado a un nivel de HU. Los tres tipos de ventanas elegidas son:

Brain Window o ventana del cerebro, cuyos valores de ventana son centro: 40 HU y ancho: 80 HU es capaz de extraer las estructuras propias del cerebro.

Subdural Window o ventana subdural, cuyos valores de ventana son centro: 80 HU y ancho: 200 HU con el fin de hacer más visible la ICH. Ayuda en la detección de los delgados hematomas subdurales agudos, los más difíciles de detectar debido a su forma y localización.

Soft Window o ventana de tejidos blandos, cuyos valores de ventana son centro: 40 HU y ancho: 380 HU con el fin de extraer los nervios, vasos sanguíneos y diferentes membranas que puedan ayudar a la identificación de la hemorragia.

Después de aplicar cada ventana, se combinan en una sola imagen de 3 canales. Cada canal de la imagen corresponde a una de las ventanas generadas teniendo como resultado una imagen de 256x256x3 que será la entrada a la red neuronal que se explicará con detalle más adelante.

Finalmente, para aquellos modelos dónde no se utilice la técnica de Windowing se obtiene el array de pixeles directamente del fichero DICOM y se expanden los mismos valores a los otros 2 canales con el fin de establecer el tamaño de entrada que exige la red neuronal 256x256x3.

La última técnica de preprocesamiento es whitening o normalización. En función de los trabajos de Ye y Pal & Sudeep, [2][9] se considera que su aportación es totalmente evidente. Primero se obtiene la imagen de 256x256x3 fruto de la salida de la aplicación de Windowing o Sin Windowing y se reduce el valor de la intensidad de cada pixel en valores reales entre 0 y 1 resultante de restar la media y dividirlo por la desviación típica.

Arquitectura del Modelo

Como se ha comentado previamente en la fase exploratoria, el uso de redes neuronales para la resolución de este problema de clasificación es una tendencia muy adoptada en los últimos años para este tipo de problemas de clasificación con imágenes médicas [4][2][3][5][8].

Tras diversas pruebas para poder elegir la red que mejor se ajuste, se tomó la decisión de utilizar una de las redes que proporciona Keras dentro de su librería: la red neuronal ResNet 50. Esta red es una de las más utilizadas y que mejor resultado da para este tipo de problemas de clasificación. Este tipo de redes están diseñadas para trabajar con imágenes de 3 canales RGB, de ahí la necesidad en el preprocesado de que por cada TC era necesario la expansión a 3 canales. Por lo tanto, el input para la red neuronal ha sido de 256x256x3.

Para todas las pruebas, se ha aplicado la técnica de Transfer Learning haciendo fine-tuning sobre la ResNet 50, es decir, que únicamente se aprovechan los pesos preentrenados con ImageNet de las primeras capas, cuyos parámetros no se alteran quedándose fijos durante todo el entrenamiento. En cuanto a las últimas capas preentrenadas de la red se desechan y se sustituyen por tres capas: dos capas de 64 neuronas Fully Connected y una capa Softmax de 5 neuronas de salida que representan las 5 posibles clases (EPI, INP, INV, SHA y SDA).

Además, se ha añadido una regularización dropout del 30% con el fin de evitar el posible overfitting. Por último, tras una numerosas de pruebas se decidió utilizar el optimizador avanzado RMSProp con una tasa de aprendizaje o learning rate de 10-4 y la entropía cruzada categórica (categorical cross-entropy) como función de coste o pérdida.

Balanceo de Clases y Estrategia de Entrenamiento

Como se ha comentado previamente la distribución de los datos, el balanceo entre clases positivas y negativas está relativamente balanceado alrededor del 54% y el 46% respectivamente. Sin embargo, el objetivo de este trabajo no es detectar únicamente si el paciente en esa TC tiene o no ICH si no que, además, es necesario averiguar el tipo de ICH que padece. Es aquí dónde radica el problema ya que tenemos las clases de los tipos tenemos las clases resultantes muy desbalanceadas dentro del dataset. Se han propuesto 2 estrategias para poder comprobar la efectividad entre ellas:

Fijo y desequilibrado: No alterar el contenido del conjunto de entrenamiento en cada época. De esta manera, cada época procesa las 87.976 TC tal y como se aprecia en la Tabla I. Con esta estrategia mantendremos un claro desbalanceo de los datos.

Dinámico y equilibrado: Alterar el contenido del conjunto de entrenamiento en cada época de tal manera que el número de datos procesados de cada clase en cada época sea el mismo. De esta manera, cada época procesa las 6.000 TC extrayendo de manera aleatoria 1.000 TC de cada clase. Con esta estrategia evitaremos que la red se entrene en cada época con más datos de una clase que de otra.

Por otro lado, otra de las alternativas que tratan de resolver el problema del desequilibrio de la representación de las clases es la utilización de la técnica de Data Augmentation (DA) en el entrenamiento de los datos. Está técnica era una de las líneas futuras de recomendación que hizo Darías en su trabajo [4]. Con esta técnica lo que se busca es tener la capacidad de expandir el volumen de datos de nuestro dataset realizando una serie de transformaciones en los datos con el fin para enriquecer el entrenamiento del modelo con nuevas observaciones a partir de las que ya existen. Se han propuesto 3 estrategias para poder comprobar la efectividad entre ellas:

No usar DA. La red neuronal recibirá las observaciones tal y como son, sin establecer ninguna transformación en las imágenes médicas.

Uso de DA aplicado a todas las imágenes del dataset. La red neuronal obtendrá cada una de las observaciones del dataset con una transformación aplicada de tal manera que aumenta la variedad entre las mismas.

Uso de DA aplicado a únicamente a las imágenes del dataset etiquetadas como tipo de ICH Epidural. La red neuronal obtendrá cada una de las observaciones del dataset tal y como es sin aplicar ningún tipo de transformación salvo en las de tipo Epidural las cuales se aplicará una transformación de tal manera que aumenta la variedad entre las mismas consiguiendo elevar la representación de esta clase frente al resto.

Con el fin de afectar lo menos posible al objetivo principal de este trabajo, en aquellos elementos donde se aplica DA, se ha decidido utilizar la misma configuración, es decir, siempre que se aplique DA, ya sea en una sola clase o en todo el conjunto de datos, se aplican las siguientes técnicas: volteado horizontal a un 25%, volteado vertical a un 10%, recorte por cada lado de la imagen por un valor de píxeles aleatorios muestreados uniformemente en el intervalo discreto entre 0 y 25. Esta configuración sólo se aplica al 25% de las observaciones de cada batch.

Las diferentes estrategias de DA, así como los diferentes algoritmos de

entrenamiento que se acaban de presentar, son totalmente compatibles para ser combinados por lo que en este trabajo se ha definido una serie de pruebas que ayudan a comparar de mejor manera cada una de estas técnicas. A continuación, en la Tabla II se exponen cada una de las pruebas que se han decidido realizar:

TABLA II

LISTADO DE PRUEBAS

Test	Windowing	DA	Estrategia de Entrenamiento
1	No	EPI	Dinámico y Equilibrado
2	No	Todo	Dinámico y Equilibrado
4	Sí	Todo	Fijo y Desequilibrado
5	Sí	No	Fijo y Desequilibrado
6-2B	Sí	Todo	Dinámico y Equilibrado
7-1B	Sí	EPI	Dinámico y Equilibrado
8	Sí	No	Dinámico y Equilibrado

Listado de pruebas a ejecutar en este trabajo. Se identifican las características más importantes y diferenciadoras entre cada una de ellas.

Criterios de Éxito, Métricas y Estrategia de Test

Como métricas y criterios de éxito se utilizan principalmente dos métricas que aportan cómo de bueno ha sido la predicción de cada una de las clases en cada modelo o prueba: Exactitud o accuracy y F1 Score.

Todas estas métricas se tomarán a cada una de las clases con el fin de detectar cómo de bueno es el modelo para clasificar las observaciones en cada una de ellas. Por lo tanto, con el fin de obtener una comparación lo más real e igualitaria posible entre todos los modelos a probar, se han establecido varios subconjuntos de test donde se escogerá un número concreto de observaciones uniforme para cada las clases. Es decir, 300 observaciones para construir cada una de las clases que compondrán cada subconjunto y 1500 observaciones para el subconjunto total.

Por último, los resultados se contrastan con el trabajo de referencia de H. Ye [2] con el fin de comprobar cómo de bueno ha sido el entreno del modelo respecto al de H. Ye y comprobar la aportación realizada a la investigación.

V. DESARROLLO Y RESULTADOS

Todas las pruebas presentadas en el apartado anterior comparten ciertas configuraciones y técnicas comunes tal y como se expresó en dicho apartado. Estas son: el uso de redimensionado a 256x256x3 y su posterior normalización, uso de Transfer Learning sobre la ResNet 50 previamente entrenada con ImageNet donde sólo se entrenan las Fully Connected Layers del final, función de optimización: RMSProp partiendo de un Learning Rate 10^{-4} y su función de coste asignada Categorical Cross Entropy. A continuación, se detalla el desarrollo de cada una de las pruebas, así como sus objetivos y resultados.

A. Evaluación 1 (Pruebas 1 y 2)

Las dos primeras pruebas han sido planteadas con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento frente a cuando se aplica únicamente a la clase con menor representación (Epidural) sobre un escenario dónde no se aplica la técnica de Windowing como preprocesamiento. Ambas pruebas, además comparten la misma estrategia de entrenamiento a la hora de distribuir las clases de una manera dinámica y equilibrada en cada una de las épocas del training.

En la primera prueba, se ha entrenado el modelo con 1000 elementos de cada clase en cada época, haciendo una selección aleatoria de los 1000 primeros elementos de cada clase. La clase Epidural sólo dispone de 1047 elementos en todo el conjunto de entrenamiento, esto significa que en la siguiente la siguiente época se volvería a procesar al menos 953 observaciones repetidas de la clase Epidural cuando para otras clases lo más probable es que procesen nuevas observaciones. Por lo tanto, por la sospecha que existe de que el modelo pueda memorizar los elementos de esta clase llegando a no generalizar por overfitting, se decide aplicar DA sólo para esta clase.

Por otro lado, en la segunda prueba, se ha entrenado de igual manera el modelo con 1000 elementos de cada clase en cada época, haciendo una selección aleatoria de los 1000 primeros elementos de cada clase. En este caso, pensando que a medida que transcurriera el entrenamiento podría producirse un agravio en las clases que no tienen aplicado DA, se decide comprobar, qué resultado se obtiene en el caso de que en todas las clases se aplique la misma técnica de DA.

El entrenamiento de ambos modelos fue similar. Se estuvieron entrenando hasta que ver que la red no proporcionaba mejores resultados. Los resultados de ambas pruebas fueron bastante pobres como se puede apreciar en la Tabla III. Existieron síntomas de overfitting sin tendencia de mejora.

TABLA III

RESULTADO PRUEBA 1 Y PRUEBA 2

Clase	P1 Accuracy / F1 Score		P2 Accuracy / F1 Score	
NEG	0,66	-	0,66	-
EPI	0,70	-	0,70	-
INP	0,66	-	0,66	-
INV	0,66	-	0,66	-
SAH	0,66	-	0,66	-
SDA	0,66	-	0,66	-
Total	**0,72**	**0,05%**	0,72	**0,03%**

Resultado de las pruebas P1 y P2 Accuracy y F1 Score (sólo subconjunto total debido al pobre resultado) tras 200 épocas de entrenamiento.

B. Evaluación 2 (Pruebas 4 y 5)

Esta nueva pareja de pruebas ha sido planteada con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento frente a no aplicarlo sobre un escenario donde se usa la técnica de Windowing como preprocesamiento extra. En este caso, además se comparte la misma estrategia de entrenamiento por ambos modelos entrenándolos con un conjunto de training desequilibrado siendo las mismas observaciones en cada una de las épocas con el matiz de que en el caso de la prueba 4 estas observaciones serán recibidas por la red neuronal con ligeras transformaciones (DA) que marcan la diferencia frente a la prueba 5.

En ambas pruebas, se han entrenado los modelos con aproximadamente 87.000 observaciones con una distribución

desequilibrada de representación de las clases. En la prueba 4, se aplica DA con el fin añadir diversidad a los datos de entrenamiento y ver qué resultado tiene respecto a no utilizarlo como se ha realizado en la prueba 5.

El entrenamiento de ambos modelos fue relativamente similar. En ambas pruebas, aunque no se detectó síntomas de overfitting (como en las pruebas 1 y 2), se estuvieron entrenando hasta que ver que la red ofrecía un progreso de una manera muy lenta. En la prueba 4, el entrenamiento llegó a las 53 épocas lo cual supuso 80 horas de entrenamiento. Por otro lado, en la prueba 5, el entrenamiento llegó a las 95 épocas lo cual supuso 120 horas de entrenamiento. En comparación con los resultados obtenidos en las pruebas 1 y 2, los resultados han sido bastante mejores con síntomas de avance, sin embargo, este avance se ha producido de una manera muy lenta.

TABLA IV

RESULTADO PRUEBA 4 Y PRUEBA 5

Clase	P4 Accuracy	P4 F1 Score	P5 Accuracy	P5 F1 Score
NEG	0,96	-	0,95	-
EPI	0,71	0,0	0,70	0,01
INP	0,77	0,27	0,95	0,92
INV	0,79	0,30	0,93	0,88
SAH	0,76	0,25	0,92	0,86
SDA	0,93	0,85	0,97	0,96
Total	**0,81**	**0,30**	**0,91**	0,**66**

Resultado de las pruebas P4 (época 53) y P5 (época 93) Accuracy y F1 Score.

C. Evaluación 3 (Pruebas 6-2B y 7-1B)

Este último conjunto de pruebas ha sido planteado con el objetivo de comprobar qué resultado de predicción se obtiene de los modelos cuando se aplica DA a todo el conjunto de entrenamiento, cuando se aplica DA sólo a la clase con menor representación y cuando no se aplica DA, todo ello sobre un escenario donde se aplica la técnica de Windowing como procesamiento extra y, además, comparten la misma estrategia de entrenamiento a la hora de distribuir las clases de una manera dinámica y equilibrada en cada una de las épocas del training.

En definitiva, haciendo hincapié sólo en estas tres pruebas se puede ver que las tres comparten todo menos la manera en la que se aplica DA. Por un lado, es interesante ver cómo se comporta DA sobre este escenario, pero lo más interesante es la comparación de los resultados de estas pruebas con los de las pruebas citadas anteriormente con el fin de confirmar si existe mejora o no en cada una de estas condiciones.

La estrategia de distribución de las clases para el entrenamiento es similar a la de las pruebas 1 y 2 y únicamente difieren en el uso de técnica de preprocesado Windowing. Siendo más concretos, la prueba 6-2B es análoga la prueba 2 con única diferencia del uso de Windowing en la presente y, en la prueba 7-1B, sucede lo mismo respecto a la prueba 1.

Por otro lado, las técnicas de preprocesado usadas en las pruebas 4 y 5 son similares, pero difieren en la estrategia de distribución de clases para entrenar. El entrenamiento de los tres modelos fue relativamente similar ya que los resultados en todas ellas han sido totalmente satisfactorios. En todas ellas, aunque no se detectó síntomas de overfitting, se estuvieron entrenando hasta que ver que la red ofrecía un progreso de una manera demasiado lenta. En la prueba 6-2B, el entrenamiento llegó a las 700 épocas lo cual supuso 84 horas de entrenamiento. Por otro lado, en las pruebas 7-1B y 8, el entrenamiento llegó a las 900 épocas lo cual supuso 90 horas de entrenamiento por cada modelo.

En comparación con los resultados obtenidos en el resto de las pruebas anteriores, estos han sido definitivamente los mejores de todo el estudio. A continuación, se comparan los mejores resultados obtenidos en cada modelo:

TABLA V

RESULTADO PRUEBA 6-2B Y PRUEBA 7-1B Y PRUEBA 8

Clase	P6-2B Acc / F1 Score		P7-1B Accuracy / F1 Score		P8 Accuracy / F1 Score	
NEG	0,91	-	0,94	-	0,93	-
EPI	0,97	0,96	0,98	0,94	0,98	0,98
INP	0,94	0,87	0,96	0,94	0,97	0,95
INV	0,96	0,93	0,96	0,93	0,96	0,94
SAH	0,94	0,87	0,97	0,95	0,96	0,94
SDA	0,94	0,89	0,96	0,90	0,97	0,96
Total	**0,95**	**0,84**	**0,95**	**0,90**	**0,96**	**0,91**

Resultado de las pruebas P6-2B P7-1B y P8 Accuracy y F1 Score tras 502, 806 y 803 épocas de entrenamiento respectivamente. Son los mejores resultados obtenidos por cada modelo.

VI. DISCUSIÓN O ANÁLISIS DE RESULTADOS

En relación con las pruebas 1 y 2, que trataron de comparar el resultado de aplicar DA de diferente manera bajo un escenario donde no se aplicaba el preprocesado de ventanas (Windowing) y la distribución del conjunto de entrenamiento se hizo de manera dinámica y equilibrada, como se pudo apreciar, no hubo apenas diferencia en sus resultados. Cabe destacar la métrica del F1 Score que fue sensiblemente más baja en el modelo donde se aplica DA a todo el conjunto de datos de entrenamiento (prueba 2) respecto al que sólo aplicaba DA a la clase con menor representación (prueba 1). Esta diferencia fue sólo de un 0.02%.

Llama la atención esta pequeña diferencia de resultados, donde los primeros indicios parece que apuntan a que el uso de DA es sensiblemente contraproducente en este escenario.

Al termino de estas dos pruebas, se concluyó que ambos resultados fueron inesperadamente muy bajos por lo que obligó a seguir continuando en la búsqueda de otras combinaciones con el fin de obtener más pruebas con las que completar esta comparativa esperando encontrar mejores resultados.

Por lo tanto, la siguiente tanda de pruebas, las pruebas 4 y 5, trataron de comprobar la efectividad de DA, pero en un escenario donde se aplicaba el preprocesado de ventanas a las TC y otras estrategias de distribución de las observaciones para el entrenamiento diferentes de las que se aplicaron en las dos primeras pruebas.

En estas dos pruebas se puede apreciar una mejora notable de rendimiento respecto a las pruebas iniciales. El uso de la técnica de preprocesado de ventanas junto con la distribución de los datos de entrenamiento de manera fija y equilibrada han conseguido aumentar las métricas de casi todas las clases respecto a las que se obtuvieron en las pruebas 1 y 2.

Tal y como se comentó en apartados anteriores, existe un claro ejemplo donde se demuestra la importancia de no fiarse únicamente de la métrica de la exactitud o accuracy. En la prueba 4, más en concreto, en los resultados del subconjunto de test para la clasificación Epidural, la

exactitud ronda el 75% y, sin embargo, el F1 Score es del 0%. Esto significa que el modelo es incapaz de clasificar adecuadamente las hemorragias etiquetadas como Epidural pero sí que ha sido capaz de identificar correctamente que los otros subtipos de ICH no estaban presentes realmente, de ahí el porqué del 75% de exactitud. En este caso, para los 300 casos etiquetados con ICH Epidural de este subconjunto, el modelo no ha clasificado a ninguno como ICH de subtipo Epidural, lo cual está muy alejado del objetivo de este trabajo. En cuanto al resto de subtipos han sido clasificados con un F1 Score entre un 71% el más bajo y un 93% el que más representación tiene, obteniendo un 30% en el Total.

Por otro lado, en el caso de la prueba 5, a pesar de que ésta tiene mejores resultados en comparación con la prueba 4, también se puede ver el mismo problema a la hora de clasificar la ICH Epidural (F1 Score 0.01%). En cuanto al resto de resultados se aprecia una mejora muy notable en exactitud y F1 Score de la mayoría de las clases positivas, superando el 90% y rondando el 89% respectivamente, obteniendo un 66% en el Total.

Por lo tanto, es en la clase Epidural donde ambos modelos suspenden. Esto es debido a la poca representación que existe entre la clase epidural y las restantes. En cada epoch de 87.000 observaciones, cuando de la clase epidural se han procesado 1.047 observaciones, de la clase Subdural se han visto más de 20.000. Esto hace que los pesos de la red se vean más afectados de cara a detectar ICH de clases con mayor representación por lo que la red no llega a ser buena generalizando las ICH de tipo Epidural y bastante buena a la hora de clasificar ICH de tipo Subdural.

Como se puede apreciar en los resultados de estas dos pruebas, la diferencia de estos ya empezó a ser notable. En este nuevo escenario, se pudo ver que el uso de DA (prueba 4) parecía que tenía un efecto bastante negativo respecto a si no se usaba (prueba 5), una diferencia del F1 Score Total de un 36% entre ambas, la cual era muchísimo mayor del que se tuvo con las 2 primeras pruebas (0.02%). Sorprendentemente, incluso utilizando técnicas de DA, que se podría llegar a pensar que (a priori) aportan mejores resultados utilizando DA que si no, se vuelve a comprobar que el resultado es peor usando DA para todo el conjunto que usándolo para sólo una clase o no usándolo directamente.

Todo esto podría verse explicado debido a que DA, sobre todo, funciona muy bien cuando el conjunto de datos es pequeño y en este caso, 87.000 observaciones no lo es tanto. Por otro lado, también se puede llegar a pensar que las transformaciones que DA ha realizado son tan exageradas que podría ser que el modelo las interpretase como observaciones totalmente dispares y le cueste llegar a aprender todos los mismos, e incluso, necesitando un tiempo mayor de entrenamiento del que se ha dedicado.

Por último, tratando de comparar el resultado de la prueba 4 con los resultados de la prueba 2, donde ambas usaban técnicas de DA aplicadas a todas las clases, se puede apreciar que hay dos diferencias notables: la estrategia de la distribución de las clases en el conjunto de entrenamiento y el uso o no de Windowing. Debido a heterogeneidad de ambas técnicas, ha sido necesario seguir investigando con el fin de encontrar cual es la pieza clave dentro de este rompecabezas para encontrar la mejor predicción en la detección de ICH y la correcta clasificación de sus subtipos. Llama la atención ver que a pesar de la cantidad de tiempo de entrenamiento y del gran volumen de los datos aportado gracias a DA, el modelo haya conseguido, sobre todo en las pruebas 1 y 2, unos resultados tan bajos.

Como conclusión y analizando los resultados vistos hasta ahora, da la sensación de que la técnica de DA no consigue aportar tanto cómo se esperaba. Por otro lado, la estrategia de distribución de las clases de un conjunto de entrenamiento fijo y sobre todo desbalanceado parece que ha aportado mejores resultados, pero, como se puede apreciar, aquellas clases con menor representación tienen una penalización muy elevada que obliga a seguir investigando. Probablemente, el motivo de estos resultados vistos hasta ahora radiquen en la incapacidad que tiene la red de poder identificar los detalles de la imagen relacionados con la ICH y la anatomía del cerebro debido a la carencia de técnicas de preprocesado de ventanas que permitirían identificarlos.

Por ello, con la siguiente tanda de pruebas, pruebas 6-2B, 7-1B y 8, trataron de sacar de dudas esta última hipótesis, repitiendo, por un lado, la prueba 1 y 2 pero usando la técnica de preprocesado de ventanas. Además, al ver el pésimo efecto de DA hasta ese momento, se decidió añadir una tercera prueba, similar a la 6-2B y 7-1B pero sin el uso de esta técnica.

Satisfactoriamente, el resultado de estos tres modelos fue un éxito respecto a los resultados obtenidos en pruebas anteriores. Además, la diferencia de resultados entre las tres pruebas ya no es tan evidente. Únicamente hay una diferencia moderada entre la prueba 6-2B, que es la que aplica DA a todas las clases del modelo, y las otras dos que, o no aplica DA, o sólo la aplican en la clase con menor representación. Con esto, se vuelve a reafirmar que para este tipo de problemas DA no aporta gran valor, incluso, todo lo contrario, es contraproducente usarlo.

En cuanto a las pruebas 7-1B (DA en Epidural) y 8 (Sin DA), la diferencia es prácticamente nula, siendo igual o sensiblemente mejor, con un incremento de 0.01% de F1 Score en prácticamente todas clases, el modelo que no presenta uso de DA.

De esta manera, se ha comprobado la importancia del uso de las técnicas de preprocesado de ventanas junto con el redimensionado a 256x256 y la normalización, además del cuidado que hay que tener a la hora de elegir una buena estrategia para servir los datos a la red de una manera equilibrada y dinámica y evitando en todos los casos el uso de DA. Con todo esto, se ha conseguido demostrar el efecto que tiene el usar o no este tipo de técnicas.

Por último, como se comentó en apartados anteriores, se compara los resultados obtenidos del mejor modelo, prueba 8, frente los resultados recogidos del modelo de referencia trabajo de H. Ye [2] sobre el que se decidió comparar los resultados. Se puede apreciar que en todos los subtipos de ICH se consigue mejorar notablemente las métricas de la exactitud y del F1 Score:

TABLA VI

COMPARATIVA DE RESULTADOS

Clase	Ye 2019 Acc / F1 Score [2]		P8 Accuracy / F1 Score		Diff Accuracy / F1 Score	
EPI	0,96	0,72	0,98	0,98	+0,02	+0,26
INP	0,90	0,93	0,97	0,95	+0,07	+0,02
INV	0,91	0,87	0,96	0,94	+0,05	+0,07
SAH	0,83	0,78	0,96	0,94	+0,13	+0,16
SDA	0,94	0,84	0,97	0,96	+0,03	+0,12

Comparativa de resultados entre el modelo presentado en el trabajo de H. Ye [2] y el mejor modelo presentado en este trabajo.

VII. CONCLUSIONES

A continuación, se reúnen las conclusiones derivadas de este Trabajo de Fin de Master en base a los objetivos específicos previamente mencionados en el apartado III.

Los primeros objetivos consistían en la revisión sistemática de la literatura actual, esta tarea se abordó de forma minuciosa en el apartado II, concluyendo de que había una falta de claridad en la efectividad del uso de algunas técnicas de preprocesado de datos y métodos de balanceo de estos por lo que se planteó una comparativa para ver cómo

de efectivos eran este tipo de técnicas en el ámbito de la detección de ICH y sus subtipos. En base a ciertos trabajos de relevancia que se tuvieron en consideración, se decidió poner en evaluación la efectividad las técnicas de preprocesado de Windowing y diferentes métodos de balanceo de datos como el uso Data Augmentation para la resolución del citado problema.

De esta manera, tal y como se indicó en el objetivo número 3, se definió un banco de pruebas donde se trataba de comprobar la efectividad de estas. Siguiendo los pasos del objetivo 4, tras el entrenamiento de los modelos y la posterior evaluación de cada uno con los mismos conjuntos de tests, se obtuvieron los resultados de cada una de las pruebas.

En base al último objetivo y teniendo en cuenta estos resultados, se realizó una comparación de estos y se pudo demostrar que el uso de la técnica de preprocesado de ventanas junto con la manera en que se alimenta la red a la hora de entrenar el modelo con unos datos distribuidos de una manera equilibrada y dinámica en cuanto al grado de representación de las clases, son la pieza fundamental en la capacidad de generalización y una elevada predicción del modelo. Además, también se demostró que el uso de técnicas de balanceo de datos generando datos sintéticos (Data Augmentation) como entrada al modelo, tenía un efecto contraproducente por lo que se ha desaconsejado el uso del mismo.

Por lo tanto, las opciones que se acaban de comentar son las técnicas y métodos recomendados a la hora de implementar modelos de diagnóstico de ICH y la clasificación de sus subtipos. No obstante, debido a la naturaleza de los datos y la versatilidad de estas técnicas y a la sencillez de su utilización, también se puede utilizar para la detección de otro tipo de enfermedades que pudiera detectar un radiólogo por medio de Tomografías Computarizas, tales como cancer, alzheimer, ulceras, lesiones de cualquier tipo, otro tipo de hemorragias, etc. … ya que no difieren notablemente en los procedimientos que se tienen que seguir para su diagnóstico.

Todo lo comentado en esta conclusión, así como, durante toda la fase comparativa experimental, responde a la pregunta planteada durante todo el trabajo de ¿qué técnicas, de las opciones disponibles, logra los mejores resultados?

En cuanto a las contribuciones aportadas por este trabajo se pueden resumir de la siguiente manera:

Se ha analizado y resumido de una manera detallada todo el estado del arte del diagnóstico de las hemorragias intracraneales y la identificación de sus subtipos a través de tomografías computarizadas a nivel clínico y el diagnóstico automático a través de algoritmos de inteligencia artificial.

Se ha proporcionado las recomendaciones necesarias para aplicar a los modelos de detección automática en relación con las técnicas de preprocesado y métodos de balanceo de datos que se deben utilizar o no.

Se ha demostrado a través de uno de los modelos evaluados que es posible mejorar los resultados de predicción de trabajos recientes en relación con el diagnóstico automático de las ICH y la clasificación de los subtipos.

Es importante concluir con que no se puede afirmar rotundamente de que este tipo de modelos se puedan llevar a producción y que actúen de manera autónoma fácilmente. A pesar de haber obtenido unos resultados bastante destacables, incluso en comparación con los últimos trabajos publicados sobre este campo, los modelos aún tienen un largo camino de mejora hasta llegar a un 100% de exactitud.

No obstante, el rendimiento de este tipo de modelos podría comprobarse de otra manera. Partiendo de la idea de que el uso de Inteligencia Artificial en medicina busca reducir el error humano, se podría valorar cómo de bueno es un modelo a partir de una comparación directa con radiólogos. En este escenario, un modelo podría no necesitar resultados prácticamente perfectos, sino simplemente igualar o superar a los radiólogos. Además, su uso se plantearía como un complemento al trabajo de los radiólogos, y no como un sustituto, permitiendo al profesional tener siempre la última palabra, pero permitiéndole reducir sus errores de forma notable. Así, la implementación de estos modelos por parte de profesionales sin conocimiento profundo de la enfermedad podría ser más razonable, aunque siempre sería necesario un procedimiento de evaluación basado en la comparación directa con radiólogos.

Por esto y debido a la escasez de tiempo que se ha permitido invertir en este Trabajo de Fin de Master, se plantean las diferentes líneas de futuro que podrían ayudar a mejorar los resultados de modelos de este tipo.

En cuanto al balanceo de datos, se podría usar técnicas de estratificación de los datos con el fin de realizar un undersampling de una manera más cuidadosa eliminando observaciones que no fueran tan representativas y evitando hacerlo con aquellas que tuvieran más valor.

En cuanto al preprocesamiento de los datos, sería interesante hacer algún tipo de segmentación de las imágenes con el fin de extraer únicamente el cráneo, de tal manera que se reduciría notablemente el esfuerzo computacional a la hora de entrenar el modelo.

Por otro lado, sería también muy interesante cambiar la última capa softmax del modelo por una capa multi-class que nos permitiera dar una distribución de probabilidad de cada una de las clases con el fin de clasificar correctamente a los pacientes que tuvieran más de un tipo de ICH.

Por último, también se debe tener en cuenta los avances en el mundo del Deep Learning en general con el fin de abrir nuevas vías de investigación para esta línea de investigación. El mundo de la Inteligencia Artificia avanza muy rápido y cada día se publican nuevas técnicas o arquitecturas que mejoran enormemente a las anteriores, y que permiten el entrenamiento de redes más complejas, capaces de conseguir una convergencia mucho más rápida, con mejores resultados, ...

REFERENCIAS

[1] Díaz-Guzmán J, B.-P. F.-L. (2008). Prevalence of Stroke and Transient Ischemic Attack in Three Elderly Populations of Central Spain. Neuroepidemiology, 30(4), 247-53.

[2] Ye, H. G. (2019). Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. European radiology, 29(11), 6191-6201.

[3] Arbabshirani, M. R. (2018). Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ digital medicine, 1(1), 1-7.

[4] Darias Plasencia, Ó. (2019). Medicina personalizada: comparativa de técnicas para el diagnóstico automático del Alzheimer. Logroño (España): Universidad Internacional de la Rioja.

[5] Esteva, A. (2017). Dermatologist-level classification of skin cancer. Nature, 542(7639), 115.

[6] Gulshan, V. (2016). Development and validation of a deep learning . JAMA - J. Am. Med. Assoc., 2402-2410.

[7] Maciej A. Mazurowskia, P. A. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks, 427-436.

[8] Ding, Y. S. (2018). A Deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain. Radiology, 290(2), 456-464.

[9] Sudeep, K. K. (2017). Preprocessing for image classification by convolutional neural networks. IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). doi:10.1109/RTEICT.2016.7808140