Aplicación de redes neuronales convolucionadas de una dimensión para estudiar la respuesta a la terapia en glioblastomas

Paula Daniella Restrepo Galvis
Orlando Hernández Rodríguez

Trabajo de grado para optar al título de
Magíster en Ciencia de Datos

Directora
Profesora Sandra Ortega-Martorell
Ciencia de la computación, PhD

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería Industrial
Decanatura de Ingeniería de Sistemas
Decanatura de Matemáticas
Maestría en Ciencia de Datos
Bogotá D.C., Colombia
2022

Publicado en 2023 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Reconocimiento o Agradecimientos

Agradecemos a nuestras familias por el apoyo constante en el emprendimiento de nuestros proyectos académicos y la confianza depositada en nosotros para seguir creciendo profesionalmente.

A la profesora Sandra Ortega-Martorell, una de las primeras docentes que tuvimos en la maestría y quién inmediatamente nos transmitió su pasión por los temas relacionados con la ciencia de datos. Nos sentimos orgullosos de haber contado con una formadora íntegra y experta en la academia e investigación. Su gran trayectoria se ve reflejada en la mejora continua por desarrollar espacios para que el aprendizaje fuera de alta calidad y productividad. Sin su empatía y carisma por sus estudiantes, el resultado y el proceso de esta maestría no hubiese sido el mismo. Esperamos responder a su confianza en nuestro trabajo, y reflejar nuestros aprendizajes en la presente investigación.

A la profesora Sonia Jaimes por su iniciativa al cooperar en la creación de esta maestría en nuestro Alma Mater.

A los demás profesores, quienes proactivamente buscaron la mejor forma para transmitir su conocimiento y profundizar en nuestras inquietudes. Valoramos los aportes de los docentes y compañeros con los que aprendimos colectivamente en el desarrollo de la Maestría de Ciencia de Datos.

Resumen

Este estudio presenta la comparación entre varios modelos de aprendizaje automático con el fin de establecer el mejor modelo para predecir si hay respuesta a la terapia en pacientes comprometidos con glioblastomas (GBM), utilizando señales derivadas de imágenes espectroscópicas de resonancia magnética (MRSI). La aplicación de modelos de clasificación lineales, no lineales y redes neuronales convolucionadas de una dimensión permitirían determinar la temprana identificación del nivel respuesta a la terapia, con el objetivo de personalizar los tratamientos para el GBM y así mejorar su eficacia, pues podría incrementar la tasa de supervivencia. Actualmente, los pacientes diagnosticados con GBM reciben tratamiento de quimioterapia y radioterapia, y en algunos casos incluye la resección quirúrgica del tumor. Posterior al tratamiento, puede ocurrir remisión del tumor. Nuestro aporte radica en mejorar la interpretación de los datos obtenidos en la fase de tratamiento, para ayudar a entender, de una manera no invasiva, si los tumores están en respuesta a la terapia. Para ello estudiaremos la composición química de las muestras revelando la información metabólica (biomarcadores) (Horská & Barker, 2010), para profundizar en la investigación del GBM, uno de los tumores cerebrales más agresivos y fatales en los humanos. La comparación de estos modelos y su respectiva evaluación tendrán presente métricas relacionadas con estudios médicos, analizando el desempeño de los modelos por medio de la especificidad de los resultados y así evaluar la capacidad de discriminar los falsos negativos que se traduce en la falta de la detección temprana de la enfermedad (Komori, 2022).

Abstract

This paper presents the comparison between several machine learning models to establish the best model to predict response to therapy in glioblastomas (GBM) patients, based on the analysis of signals derived from magnetic resonance spectroscopic imaging (MRSI). This analysis would allow early identification of the therapy response, enabling personalization of treatments for GBM and thus improving their efficacy, as it could increase the survival rate. The application of linear, nonlinear classification models and one-dimensional convolutional neural networks would make it possible to determine whether there is a response to the treatment provided. Currently, after patients have been diagnosed with GBM (for which MRSI can be used), treatment would include chemotherapy and radiotherapy, even surgical resection of the tumour area. Still, remission can occur. Our contribution lies in improving the interpretation of the data obtained during the therapy to understand the chemical composition of the samples revealing metabolic information (biomarkers) from the analysis of larger areas compared to previous technologies (Horská & Barker, 2010). This is essential for further investigation of GBM, one of the most aggressive and fatal tumours in humans. The comparison of these models and their respective evaluation will take into account metrics related to medical studies, analyzing the performance of the models through the specificity of the results and thus demonstrating the ability to discriminate false negatives that results in the lack of early detection of the disease (Komori, 2022).

Tabla de contenido

Lista de Figuras

Lista de Tablas

1 Introducción

1.1 Problemática (Justificación)

El Glioblastoma (GBM) es un tipo de tumor cerebral. No es uno de los tumores más comunes en los humanos, pero sí uno de los más agresivos y fatales. El tratamiento de esta enfermedad a partir de métodos convencionales como la quimioterapia y la radioterapia ha permitido que se alcance una tasa de supervivencia entre 12 y 15 meses (Delgado-Goñi et al., 2016). Uno de los retos que impone el diagnóstico y tratamiento de este tipo de tumor es debido a su localización, ya que, al encontrarse en el cerebro, cualquier biopsia o estudio realizado supone altos riesgos. De ahí la importancia de poder contar con técnicas no invasivas para diagnosticar y monitorear el progreso de la terapia en pacientes que sufren de GBM.

La alta heterogeneidad de este tipo de tumor cerebral es otro de los grandes retos en la efectividad de los tratamientos. Esta heterogeneidad se produce por la evolución clonal, toda vez que la célula original del tumor se multiplica en clones y estos presentan diferentes sensibilidades a la terapia, con la habilidad de sobrevivir y proliferar, disminuyendo así, la esperanza de vida del paciente (Parker et al., 2015).

El diagnóstico preciso es esencial para un óptimo manejo clínico de pacientes con tumores cerebrales. Cuando es accesible, la mayoría de los tumores son eliminados quirúrgicamente, pero existe un equilibrio entre quitar tanto tejido tumoral como sea posible, manteniendo las funciones cerebrales vitales, y la radioterapia se usa a menudo para tratar cualquier tejido canceroso restante. Los recientes avances en el tratamiento de los gliomas han mejorado la supervivencia, así como la supervivencia libre de progresión, de pacientes afectados por esta patología. Como actualmente los tratamientos disponibles presentan un riesgo, es importante identificar a los pacientes que se beneficiarán de tratamientos agresivos y también a aquellos pacientes a los que el tratamiento de elección debe ser del tipo conservador.

La forma más común de diagnosticar GBM es a través del uso de imágenes por resonancia magnética (MRI) y tomografía de emisión de positrones (PET), pero ninguna de estas dos opciones puede producir biomarcadores tempranos no invasivos y robustos para confirmar la efectividad del tratamiento (Delgado-Goñi et al., 2016). Por eso, se hace necesario encontrar otras opciones que permitan al médico decidir mejores alternativas para tratar la enfermedad, basado en evidencia real sobre la evolución de esta. El uso de la espectroscopia de resonancia magnética (MRSI) ofrece información muy rica sobre los diferentes metabolitos del tejido, tanto dentro como alrededor del tumor. La interpretación de los patrones de espectroscopia es una tarea que requiere entrenamiento especializado de parte de radiólogos, y se ha visto en numerosos estudios que los métodos de aprendizaje de máquina pueden servir de apoyo a los radiólogos y médicos en esta difícil tarea.

En estudios previos del GBM se ha investigado el tratamiento de la enfermedad, las barreras encontradas para que estos sean efectivos y posibles estrategias para mejorar su efectividad.

Algunos han logrado determinar que los vasos sanguíneos del cerebro poseen una mayor protección comparados con aquellos del resto del cuerpo del ser humano. Es por esta razón que la difusión, tanto de los nutrientes como de las medicinas que viajan en el torrente sanguíneo, tendrá mayor complejidad para ingresar y salir de las células del cerebro. Lo anterior, genera que los tratamientos para el GBM resulten inefectivos (Wolburg et al., 2012).

A este problema se suma la heterogeneidad de las células tumorales que poseen los GBM, pues se han observado diferencias entre las células cancerígenas del mismo tumor. Debido a que los tratamientos (quimioterapia/radioterapia) no pueden ingresar apropiadamente a las células del cerebro como consecuencia de la barrera de los vasos sanguíneos, sólo habrá algunas células tumorales que serán atacadas. Aquellas que han creado resistencia al tratamiento, se seguirán proliferando hasta alcanzar nuevamente la magnitud del tumor original, causando nuevamente la sintomatología (Tan et al., 2020).

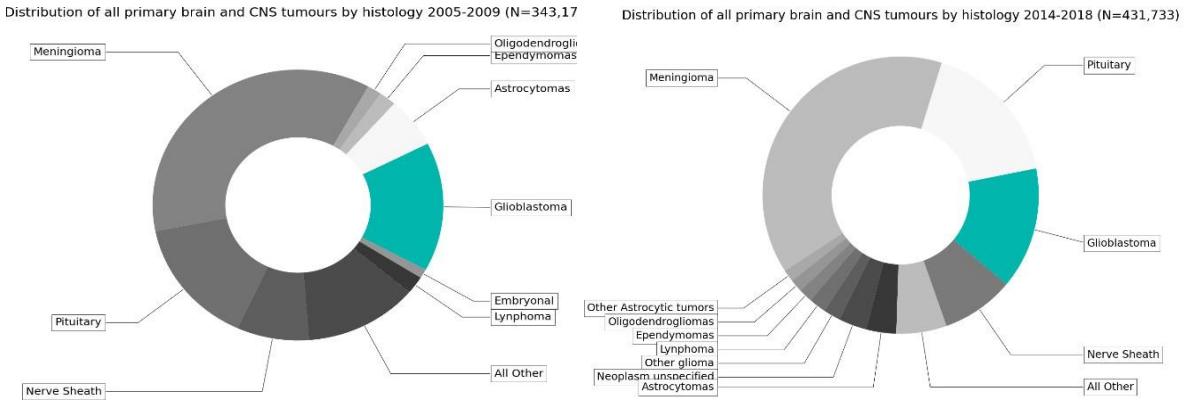
Otras barreras identificadas para el éxito de los tratamientos de GBM están relacionadas con los tipos de esta enfermedad (Classical, Proneural Mesenchymal), la evolución clonal y las células madre del cáncer.

Actualmente, se invierte en I&D enfocado en: (a) encontrar nuevas formas de identificación de células tumorales, (b) identificar los clones relacionados con las células madre del GBM, (c) identificar biomarcadores para personalizar los tratamientos, (d) desarrollar nuevas medicinas, (e) realizar seguimiento a la aplicación de inmunoterapia y (f) a la terapia con nanopartículas.

Esta propuesta de investigación profundizará en el uso de la ciencia de datos con el fin de proponer una herramienta alternativa que permita dar mejor seguimiento al tratamiento del GBM y con ello ajustar las terapias en pacientes y maximizar su beneficio.

A continuación, mostramos la tendencia de los casos confirmados para tumores cerebrales y del sistema central nervioso en estudios realizados en Estados Unidos en dos rangos de tiempo: (i) Desde 2005 a 2009 (Dolecek et al., 2012) y (ii) Desde 2014 a 2018 (Dolecek et al., 2012), clasificados histológicamente:

Figura 1 Distribución de tumores cerebrales y del SCN por histología, 2005-2009 y 2014- 2018



Fuente: Datos tomados del documento WHO Classification of Tumours 2021 (Dolecek et al., 2012).

En la Figura 1, podemos observar que el GBM ocupó un porcentaje del 15.8% en 2005-2009 (54.221 casos) y del 14.3% en 2014-2018 (61.737 casos), por la cual, existe una variación marginal en la proporción de la enfermedad dentro del espectro de los tumores cerebrales, pero también notamos que hay un incremento del 13.8% de los casos confirmados en un mismo periodo de tiempo debido a que en la actualidad se han confirmado un mayor número de pacientes para estudio. Aquí también radica la importancia de aportar a la literatura ya que la tecnología, especialmente ML, ha servido de catalizador para que la Organización Mundial para la Salud – OMS haya actualizado sus publicaciones con los nuevos hallazgos hechos por aprendizaje de máquina, evidenciándose que existe una subdivisión mucho más amplia de los tejidos comprometidos y esas apreciaciones no hacían parte de la literatura (Komori, 2022).

Así las cosas, se ha concluido que identificar con mayor eficacia el estado de los tejidos involucrados en este tipo de problemas, contribuye a mejorar los diagnósticos y posibles tratamientos para el GBM y los tumores cerebrales en general.

1.1.1 Estado del arte con relación a la ciencia de datos

Al día de hoy, existen una gran cantidad de estudios que evidencian el uso de Machine Learning (ML) y Deep Learning (DL) como herramientas eficaces para predecir y clasificar los tejidos comprometidos con tumores, como por ejemplo, el uso de métodos tradicionales como modelos para la clasificación de tumores usando CNN y luego, utilizando el multiumbral para extraer el tumor detectado (Tuhin et al., 2020), máquinas de soporte vectorial con mínimos cuadrados (Luts et al., 2007), clasificadores híbridos que usan Deep CNN y LuNet (Balamurugan & Gnanamanoharan, 2022), representación automática de característica complejas en imágenes (Kouli et al., 2022) ,todos ellos, abordando el mismo problema desde diferentes puntos de vista y acercamientos.

Si bien algunos de estos estudios proponen modelos para detectar tumores cerebrales y visualizarlos de forma precisa y eficaz, clasificando los tejidos con altos porcentajes de precisión, ninguno de ellos, ha planteado los modelos de DL y ML a partir de señales compuestas de biomarcadores que permiten un estudio más detallado de las características bioquímicas del tejido cerebral.

Un estudio previo (Delgado-Goñi et al., 2016), ha propuesto una solución basada en la extracción de fuentes para estudiar, de forma no invasiva, si hay o no respuesta a la terapia (o tratamiento), en este caso temozolamida. Este proyecto plantea la aplicación de una metodología alternativa con el fin de intentar mejorar el mismo problema clínico: por medio de la tecnología de análisis de datos se propone aplicar aprendizaje automático (ML) y aprendizaje profundo (DL) para desarrollar y evaluar diferentes modelos de clasificación, incluyendo métodos tradicionales y redes neuronales convolucionadas.

Este trabajo de investigación utilizará los mismos datos que fueron utilizados en el estudio previo mencionado (Delgado-Goñi et al., 2016), dado que el objetivo central es evaluar el uso de metodologías alternativas. Dichos datos son señales de espectroscopia de resonancia magnética, conocidas en inglés como Magnetic Resonance Spectroscopic Imaging (MRSI). Nuestra hipótesis es que los métodos alternativos propuestos en la investigación podrían ser modelos más eficientes en detectar dichos biomarcadores y en consecuencia, identificar el tejido tumoral del sano.

, por lo que el paciente, no se vería en la necesidad de ser expuesto a exámenes invasivos, sino más bien, evitaría el proceso de toma de biopsias ya que el cerebro es de complejo acceso debido a una serie de estructuras óseas, membranas nerviosas y líquidos que protegen el órgano (Moreno & Holodny, 2021). Adicionalmente, existe la posibilidad de mitigar riesgos relacionados con la afectación de otras funcionalidades como el habla, la visión o el movimiento, debido a que no habría contacto físico con la anatomía del cerebro.

1.2 Objetivos y Pregunta de Investigación

La pregunta de investigación que pretende resolverse es la siguiente: ***¿Es posible establecer un modelo de alto desempeño para predecir a qué tipo de clase pertenecen los tejidos del cerebro comprometidos con Glioblastomas (GBM), analizando las señales derivadas de imágenes espectroscópicas de resonancia magnética (MRSI), con la aplicación de modelos de clasificación lineales, no lineales y redes neuronales convolucionadas de una dimensión, con el fin de determinar si existe respuesta al tratamiento suministrado?***

Para ello, los objetivos específicos propuestos son los siguientes:

- Identificar los datos necesarios para el análisis de señales generadas a partir de imágenes espectroscópicas de resonancia magnética (MRSI) de los tejidos cerebrales comprometidos con GBM.
- Construir un pipeline para señales de MRSI obtenidas de los tejidos cerebrales comprometidos con GBM y la zona aledaña.
- Diseñar modelos de predicción sobre la clasificación de tejidos cerebrales comprometidos con GBM, con métricas de desempeño confiables y reproducibles.
- Visualizar los resultados de los modelos de predicción sobre la clasificación de tejidos cerebrales en mapas de clasificación para compararlos entre sí.
- Evaluar los modelos establecidos con el fin de determinar cuál es el más efectivo, confiable y preciso en la clasificación de los tejidos cerebrales comprometidos con GBM y determinar si contribuye al tratamiento de la enfermedad.

1.3 Alcance y Limitaciones

Con esta investigación se espera encontrar un modelo que tenga la capacidad de discernir si una zona tumoral en el cerebro está respondiendo al tratamiento o no. De esta manera, con las señales derivadas del MRSI, se busca, además de la clasificación de las características inherentes de la zona, poder realizar un mapa de clasificación que represente el área patológica del cerebro y confrontarla con las imágenes MRSI, evaluando la precisión de la predicción resultante.

Los estudios previos de GBM han contado con diferentes modelos de investigación. Algunos, han tratado de replicar las células tumorales en los laboratorios; otros, han analizado bases de datos de los pacientes, recopiladas por entidades privadas para el posterior estudio de la enfermedad y la efectividad de sus tratamientos; y otros, han hecho estudios replicando el GBM en el cerebro de animales como los ratones.

Para el desarrollo de nuestra investigación se optará por esta última opción, para lo cual se cuenta con las bases de datos generada a partir de los estudios previos realizados por (Ortega-Martorell et al., 2019) en los años 2012, 2013 y 2016, reseñados en la bibliografía de esta propuesta.

Estos datos se manipularán y analizarán por medio de la herramienta TensorFlow 2(M. Abadi, 2016). Este software de código abierto de extremo a extremo permite de forma gratuita realizar todo el análisis de la investigación, facilitando la creación de modelos de aprendizaje automático ya que combina habilidades claves para la programación por medio de infraestructura de capas. Adicionalmente, cuenta con Keras como una API de alto nivel para resolver problemas de aprendizaje automático con un enfoque especial en aprendizaje profundo moderno.

Es pertinente indicar que esta propuesta de investigación hace parte de la continuación de los estudios ya realizados por otros expertos en medicina y ciencia de la computación. La innovación de este estudio pretende aportar con una metodología que enfoque los recursos en el tratamiento del GBM con el fin de mejorar la expectativa de vida de los pacientes y sus condiciones durante la fase de la enfermedad, pues actualmente, no se cuenta con una cura efectiva, ni un procedimiento quirúrgico que permita la extracción completa del tumor.

Nuestro estudio se enfocará en la clasificación de los tejidos del cerebro a partir del análisis de señales derivadas de MRSI de los cuales se conocen las áreas de tejido Normal (N), tumor en respuesta a la terapia (R), y tumor sin respuesta a la terapia (T). La clase T también representa a los tumores control, o sea aquellos que no han recibido terapia. Estas clases, N, R y T, nos permitirán estudiar el problema con algoritmos de aprendizaje automático supervisado.

Para esto, se utilizarán tres estrategias, métodos lineales, métodos no lineales y por último redes neuronales convolucionadas de una dimensión (1D-CNN) para el análisis de los espectros (señales) que provienen de MRSI. Luego, se evaluarán los modelos y se compararán bajo el uso

de múltiples métricas de desempeño. El modelo óptimo se utilizará para predecir a qué tipo de clase pertenecen los tejidos del cerebro comprometidos con GBM.

Por otra parte, las limitaciones de esta investigación pueden estar relacionadas con:

- Tamaño de la muestra de las clases R y T, pues la última presenta la menor cantidad de registros y entre estas dos clases se presenta similitud en la estructura de las señales. Lo anterior podría presentar un desafío para el aprendizaje por parte de los modelos de predicción.
- Falta de datos disponibles debido a que los datos fuente de esta investigación son difíciles de obtener, pues al tratarse de experimentos con seres vivos (ratones) requiere de unos equipos, permisos, presupuestos y metodologías especiales. En este sentido, la investigación debió realizarse con los datos disponibles, y con ellos determinar el alcance de clasificación de los modelos, y ver si es necesario hacer nuevamente el análisis con la producción de más datos.
- El acceso a los datos es limitado, pues no se tiene acceso directo a los mismos, ya que se trata de la información científica recopilada por el Departamento de Bioquímica y Biología Molecular y la Unidad de Patología Murina y Comparada del Departamento de Medicina y Cirugía Animal, de la Universidad Autónoma de Barcelona y fueron facilitados única y exclusivamente para el desarrollo de esta investigación, sin opción de pedir ajustes o complementos a la data entregada.

1.4 Metodología

Esta investigación se desarrolla de acuerdo con la metodología Knowledge Discovery in Databases – KDD. En este sentido, la figura 2 describe se seguirá el esquema de trabajo que se seguirá:

Figura 2 Metodología KDD

1.4.1 Selección de datos

Los datos fueron seleccionados a partir de la información científica recopilada por el Departamento de Bioquímica y Biología Molecular y la Unidad de Patología Murina y Comparada del Departamento de Medicina y Cirugía Animal, ambos de la Universidad

Autónoma de Barcelona en España. A partir de los estudios realizados con el uso de cepas especiales de ratones para estudiar el GBM, se han venido realizando investigaciones con el fin de prevenir y tratar esta enfermedad.

Estos datos hacen parte de un estudio previo de mejora de los tratamientos de GBM referenciados previamente como el estado del arte de esta investigación y fueron entregados por la Directora de Investigación, la doctora Sandra Ortega-Martorell, quien hace parte de la red de investigadores de dicha Universidad.

En el desarrollo de esta propuesta de investigación se tomaron los datos relacionados con el estudio de señales que proveen información de la actividad celular (metabólica). La data de cada individuo a su vez comprende una figura compuesta por una grilla 10x10, en donde cada una de esas celdas es conocida como vóxel, que son unas unidades en forma de cuadrado con un ancho y un largo a la que conoceremos como posición. Cada uno de estos vóxeles dejarán una huella que entenderemos como señal, la cual se compone por 692 características para cada uno de 33 individuos del estudio. Estos datos corresponden al estudio pre-clínico en ratones del tipo GL261 que han sido inoculados con GBM. Dichos ratones fueron escaneados en un escáner de resonancia magnética a 7 Teslas con la técnica PRESS-MRSI utilizando tiempo de eco corto (12 ms) y largo (136 ms).

Se cuenta con dos ficheros principales, uno con la señal que proviene de cada vóxel y otro con la posición de este en la MRSI. Cada señal cuenta con 692 características que conforman su comportamiento, las cuales serán utilizadas para el entrenamiento de los modelos y, la posición cuenta con dos variables que servirán para la definición y delimitación de las zonas sanas y tumorales. Cada señal se encuentra etiquetada entre las clases X, N, R y T.

Esta selección de los datos es la parte inicial para el cumplimiento del objetivo específico 1 de esta investigación: Identificar los datos necesarios para el análisis de señales generadas a partir de imágenes espectroscópicas de resonancia magnética (MRSI) de los tejidos cerebrales comprometidos con GBM.

1.4.2 Preprocesamiento de información

De acuerdo con la información disponible de los individuos de estudio se organizaron los datos así:

- **Grupo A**: 7 individuos en estudio afectados por GBM pero que no se les ha suministrado tratamiento alguno. Este es el grupo de control de la investigación.
- **Grupo B**: 14 individuos afectados por GBM a los cuales se les ha suministrado tratamiento.
- **Grupo C**: 12 individuos afectados con GBM a los cuales se les ha suministrado tratamiento y se ha monitoreado su respuesta en distintos días.

En total, se obtienen 158 eventos derivados de señales a partir de MRSI y cada uno de ellos posee archivos separados para las observaciones con etiquetas que clasifican los tejidos de las zonas cerebrales en estudio, en:

- Normal (N)
- Tumor en respuesta a la terapia (R)
- Tumor sin respuesta a la terapia (T)
- Zonas sin definir (X).

Para cada uno de estos eventos se ha monitoreado la señal conformada por 692 variables para cada uno de los vóxeles – grilla de 10x10 para un total de 100 señales por individuo.

Adicionalmente, se ha referenciado la coordenada a la que pertenece cada señal en la grilla (vóxel). De forma preliminar, se anticipó que se trabajará con una base de datos que consta de 9.272.800 registros para entrenamiento y prueba de los modelos, lo cual permitirá estudiar el problema con algoritmos de aprendizaje automático supervisado.

Este preprocesamiento de la información resulta fundamental para la identificación de los datos necesarios para el análisis de señales generadas a partir de imágenes espectroscópicas de resonancia magnética (MRSI) de los tejidos cerebrales comprometidos con GBM, razón por la cual también hace parte del cumplimiento del objetivo 1 de la investigación.

1.4.3 Transformación de los datos

Los archivos originales se transformarán para ser unificados en una sola base de datos que comprenda las 692 variables obtenidas por las señales de las imágenes MRSI, dos variables para las coordenadas X y Y, y una última característica que corresponderá a la etiqueta para la clasificación de los tejidos: N, R, T y X. Una vez se tuvo esta base consolidada, se procederá a modelar los algoritmos propuestos con las 692 características de la señal, sin hacer uso de los registros cuya etiqueta sea la clase (X), pues esta no permitirá reconocer patrones de los demás tejidos en estudio. Con respecto al desbalanceo entre las clases restantes (N, R y T) se considerarán técnicas de balanceo de muestra con el objetivo de evitar la presencia de sesgo sobre la clase mayoritaria que, en este caso es R.

Además, se tomó esta base de datos unificada y se le aplicó un método de selección de variables con el ánimo de minimizar la función de costos por medio de un modelo lineal. En nuestro caso, se busca seleccionar aquellas características útiles y descartar aquellas redundantes por medio de un hiperparámetro que afina la intensidad de la penalidad dentro de la función de costo.

1.4.4 Minería de datos

Durante esta etapa se identifican los siguientes métodos para el desarrollo de un modelo óptimo:

- Métodos lineales: regresión logística y máquinas de soporte vectorial.
- Métodos no lineales: Random Forest y XGBOOST.
- Método de redes neuronales convolucionadas, y de redes neuronales convolucionadas de una dimensión con el objetivo de aprovechar la forma de la señal.

Es importante en este punto analizar las ventajas y desventajas del uso de cada método teniendo en cuenta el tipo de dato que poseemos para el estudio, tal como se detalla en el capítulo tercero de esta investigación. Sin embargo, preliminarmente, se considera pertinente plantear algunos conceptos importantes para cada modelo.

➢ Modelo de regresión logística para problemas de clasificación de tejidos multiclase:

De acuerdo con IBM en su publicación ¿*"What is logistic regression?",* la regresión logística es un método estadístico que puede ser utilizado para la clasificación de un problema binario. Se basa en estimar la probabilidad de ocurrencia de un evento de acuerdo con la información derivada de unas variables independientes. Se aplica una función logit a la probabilidad de ocurrencia, es decir, a la probabilidad de éxito sobre la probabilidad de fracaso.

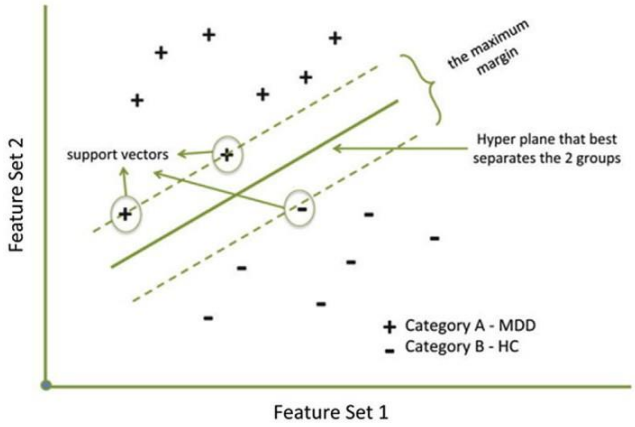
Entre las ventajas que presenta la utilización de este método de aprendizaje automático, se encuentra su simplicidad e interpretabilidad, lo que puede llegar a superar otros modelos más sofisticados. Sin embargo, puede ser sensible a la presencia de una muestra pequeña (Kirasich et al., 2018)

➢ Modelo de máquinas de soporte vectorial para problemas de clasificación de tejidos multiclase – SVM:

Este es un algoritmo de aprendizaje automático supervisado para clasificación y regresión. Una función de decisión SVM es precisamente un "hiperplano" óptimo que sirve para separar (es decir, "clasificar") las observaciones que pertenecen a una clase de otra en función de patrones de información sobre esas observaciones llamadas características. Ese hiperplano se puede usar para determinar la etiqueta más probable para los datos no vistos. Un hiperplano óptimo es aquel que maximiza el margen que existe entre clases, demostrando un mejor desempeño para discernir entre las clases del problema. Este algoritmo puede ser lineal o no lineal (Pisner & Schnyer, 2020).

Entonces, la idea principal de un algoritmo de máquinas de soporte vectorial es, si tenemos dos clases, por ejemplo, encontrar dos vectores de soporte que tengan características de la clase contraria y así construir el hiperplano que comenzará a clasificar de acuerdo con la limitación que generen estos vectores de soporte. Así como se observa en la Figura 3.

Figura 3 Definición de SVM



Fuente: Support Vector Machine (2020). Derek A. Pisner, David M. Schnyer

➢ Modelo de random forest para problemas de clasificación de tejidos multiclase

Random Forest es un algoritmo basado en una combinación de los resultados de distintos árboles de decisión, conduciendo a predicciones más confiables. La aleatoriedad de características que brinda el random forest, garantiza una baja correlación entre árboles de decisión. Su predicción se basa en el promedio de las predicciones de los árboles (Pal, 2005).

Un árbol de decisión se basa en un nodo que tendrá el mejor atributo de prueba y de acuerdo con reglas de decisión para realizar la clasificación de las características, este proceso se repite hasta efectuar la mejor segmentación (Mbaabu, 2022).

➢ Modelo de ensamble utilizando múltiples algoritmos de aprendizaje para problemas de clasificación de tejidos multiclase

Extreme Gradient Boosting (XG-Boost), es un algoritmo de aprendizaje automático del grupo de árbol de decisión, potenciado por gradiente. Es bastante utilizado actualmente debido a su eficiencia en el tratamiento de grandes cantidades de datos, distintas clases y variedad de tipos de datos. Su característica más importante es el escalamiento, potenciando el rendimiento del modelo y el uso de recursos (Chen & Guestrin, 2016).

El funcionamiento de este modelo consiste en entrenar iterativamente modelos de árbol de decisión y utilizar los residuos del error para ajustar el siguiente modelo minimizando la varianza y el sobreajuste. Al utilizar el *Gradient-boost* se minimizará el sesgo y el desajuste (Ogunleye & Wang, 2020).

➢ Modelo redes neuronales para problemas de clasificación de tejidos multiclase

Las redes convolucionadas, 2D CNN, son conocidas como arquitecturas de modelos para resolver problemas de visualización artificial (imágenes) y por eso se conoce que son utilizadas en los mejores algoritmos para la solución de problemas de cómputo visual (Verma, 2019).

- Modelo de redes neuronales convolucionadas de una dimensión para problemas de clasificación de tejido cerebral multiclase

Las redes convolucionadas de una dimensión siguen el mismo principio teórico de las 2D CNN al tratar de extraer características espaciales de la base de datos utilizando su Kernel. La gran diferencia radica en que para las 1D CNN, el Kernel se desliza en una dimensión mientras que en las 2D CNN, el Kernel lo hace en dos dimensiones (Largo y ancho) (Verma, 2019).

Continuando con el proceso de minería de datos, se procedió a realizar el ajuste de hiperparámetros con el objetivo de optimizar el desempeño de los modelos anteriormente mencionados y encontrar un modelo competitivo durante el proceso. Para esto se prueban distintas técnicas como GridSearchCV, Random Search y Bayesian Optimization.

Con respecto a las redes neuronales, se utilizaron metodologías de regularización, entre ellas tenemos: *Early Stopping*, *L1 Regularization*, *L2 Regularization*, *Dropout layer* y *Batch Normalization*. Su uso depende de la necesidad de optimización del modelo y prevención de la presencia de sobreajuste *(overfitting*) o subajuste (*underfitting*). El primero, se presenta cuando un modelo se ajusta demasiado a los datos de entrenamiento que no permite generalizar correctamente los datos de testeo, mientras que el segundo se presenta cuando el modelo no es capaz de identificar patrones y su rendimiento será insuficiente (Zhang et al., 2019) [1]

Con lo anterior se busca que el modelo se ajuste de manera óptima a los datos, proporcionando una idea de lo que está ocurriendo en la zona del cerebro y si existe una respuesta a la terapia suministrada. Por consiguiente, se pretende obtener una predicción de la clasificación del área definida por un vóxel y posteriormente dibujar un mapa que represente la zona delimitada con la información provista, de acuerdo con la posición definida en las coordenadas X y Y.

El desarrollo de esta etapa de la metodología, presentada en el tercer capítulo de este trabajo, dará cumplimiento al tercer objetivo específico de la investigación: Diseñar modelos de predicción sobre la clasificación de tejidos cerebrales comprometidos con GBM, con métricas de desempeño confiables y reproducibles.

[1] Zhang, Zhang, y Jiang, «Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems».

1.4.5 Evaluación e interpretación

Por último, en cumplimiento del quinto objetivo específico de la investigación se evalúan los modelos y se comparan bajo el uso de múltiples métricas de desempeño con el fin de determinar qué modelo prima sobre otro. Para ello, se utilizan técnicas de evaluación de modelos como precision-recall curves, sensitivity analysis (saliency maps, gradcam), además de otras mediciones relevantes para los problemas de clasificación. A partir del análisis de los resultados, se determinará el modelo de mayor efectividad. El modelo óptimo se utilizará para predecir a qué tipo de clase pertenecen los tejidos del cerebro comprometidos con GBM.

Criterios de evaluación:

Las métricas de desempeño para problemas de clasificación que se relacionarán son: exactitud, sensibilidad, especificidad, precisión y el F1-Score. Esto, con el objetivo de comprender mejor los resultados que se obtienen conforme a lo establecido en la literatura (Hossin & Sulaiman, 2015).

- La exactitud se refiere al número de predicciones correctas sobre el número total de predicciones. Lo anterior se evalúa comparando la clasificación real y los casos predichos.

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

- La sensibilidad se centra en la presencia del error tipo II. Refleja la cantidad de clasificaciones de la clase positiva que se captaron sobre la cantidad de casos positivos que debía predecir el modelo.

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

- La especificidad muestra la cantidad de casos negativos que fue capaz el modelo de clasificar como negativos o en otras palabras es conocido como la tasa de verdaderos negativos.

$$Specificity = \frac{TN}{TN + FP}$$

- La precisión equivale a la cantidad de casos de la clase positiva que fueron identificados correctamente por el modelo. Se relaciona principalmente con el error tipo I.

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

➢ El F1-Score busca encontrar el equilibrio entre la precisión y la sensibilidad como una medida armónica entre estas dos métricas. Que al igual que el Youden Index (ver capítulo 3), busca combinar dos mediciones y definir la misma importancia para cada una a la hora de evaluar la eficacia de un modelo.

$$F1 - Score = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Adicionalmente, de acuerdo con la literatura (altexsoft, 2022) se deben enunciar los términos de:

- Verdaderos positivos: refleja la cantidad de casos de la clase positiva que el modelo predijo correctamente.
- Verdaderos negativos: refleja la cantidad de casos de la clase negativa que el modelo predijo correctamente.
- Falsos positivos: se refiere a la cantidad de casos de la clase negativa que fueron predichos por el modelo incorrectamente y son clasificados como positivos, se le conoce como error tipo I.
- Falsos negativos: se refiere a la cantidad de casos de la clase positiva que fueron predichos incorrectamente por el modelo y son clasificados como negativos, se le conoce como error tipo II.

2 Consolidación y preprocesamiento de datos de señales generadas a partir de biomarcadores de MRSI de los tejidos cerebrales comprometidos con GBM

2.1 Consolidación y organización de los datos originales a partir de múltiples directorios

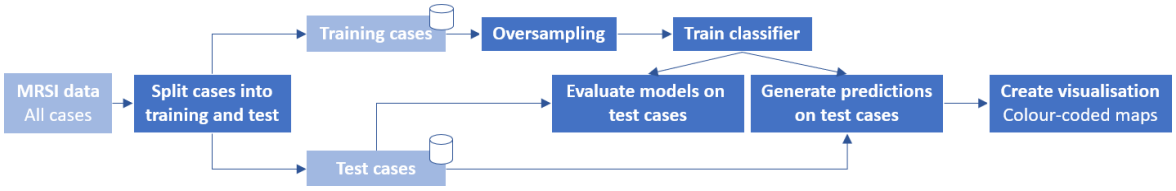
Como se indicó previamente, para el preprocesamiento de datos la información se organizó en tres grupos:

- **Grupo A**: 7 individuos en estudio afectados por GBM pero que no se les ha suministrado tratamiento alguno - Grupo de Control.
- **Grupo B**: 14 individuos afectados por GBM a los cuales se les ha suministrado tratamiento.
- **Grupo C**: 11 individuos afectados con GBM a los cuales se les ha suministrado tratamiento y se ha monitoreado la respuesta en diferentes periodos de tiempo con una sumatoria de 133 estudios.

Inicialmente, en cumplimiento del segundo objetivo específico de la investigación, se generó un pipeline. Este se define como una construcción lógica que va a representar un proceso que incluye unas fases. De esta manera se precisará el conjunto de pasos y las herramientas que se utilizarán durante el ciclo de procesamiento, extracción de características, desarrollo de los modelos, y por último, la evaluación y validación de los mismos.

En el caso de esta investigación, se generó un pipeline que permitió la generación de archivos consolidados de las señales por grupo y posteriormente se creó un archivo unificado sin distinguir la clasificación de los individuos. Este pipeline facilitó el manejo, transformación, entrenamiento y evaluación de los modelos. La arquitectura que se planteada se observa en la figura 4.

Figura 4 Arquitectura del pipeline



También se realizó un análisis para verificar la presencia de valores vacíos y atípicos, evidenciando la no presencia de estos dentro del dataset.

Con el fin de preservar la calidad de la información y evitar que exista algún factor relacionado con la data que pueda afectar el desempeño del modelo, se tomó la decisión de realizar la división del directorio en dos carpetas. La primera con datos de entrenamiento y la segunda para el testeo, de acuerdo con la proporcionalidad de las muestras para los Grupos A, B y C de los individuos analizados. En este sentido, se buscará que cada conjunto de datos se tendrá un único ratón y la información de un ratón no se encontrará en ambas carpetas, evitando que la información de testeo contenga información de un ratón que se haya usado para entrenar los modelos.

A partir del análisis exploratorio de los datos, se pudo observar que: el 84,4% de las señales se producen en los individuos del grupo C, es decir que estos registran el mayor número de eventos. El 10,4%, en el grupo B y el 5,2% en el grupo A.

Teniendo en cuenta esta información, y manteniendo estas proporciones en los datos, se seleccionaron aleatoriamente los ratones para organizar las carpetas de entrenamiento y testeo, con el objetivo de que el primer conjunto de datos contenga el 70% de la información de los eventos registrados y, el segundo posea el 30%, para cada uno de los Grupos A, B y C, respectivamente.

En la tabla 1 se puede observar la distribución de los ratones entre los conjuntos de datos de entrenamiento y testeo.

Tabla 1 Organización de los datos por individuo de estudio

Carpeta	**Grupo A**	**Grupo B**	**Grupo C**
Entrenamiento	C69, C71, C234, C278.	C288, C351, C415, C418, C525, C527, C529, C586	C821, C819, C817, C806, C797, C795, C776.
Testeo	C32, C179, C233.	C255, C437, C520, C575, C583, C584.	C809, C808, C794, C775, C774.

En total se cuenta con 158 eventos derivados de señales a partir de imágenes MRSI y cada uno de ellos posee archivos separados para el análisis, con etiquetas que clasifican los tejidos de las zonas cerebrales en estudio, en: Normal (N), tumor en respuesta a la terapia (R), tumor sin respuesta a la terapia (T) y zonas sin definir (X).

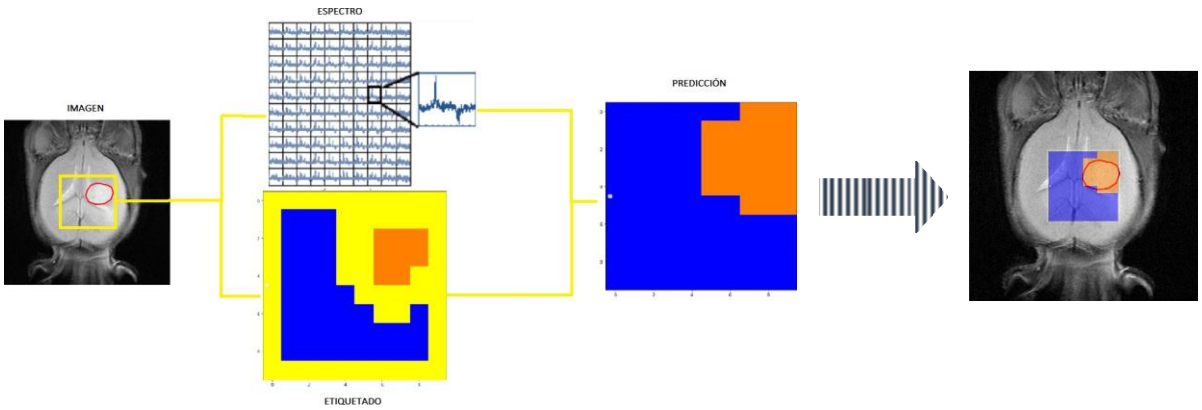
Preliminarmente, se determinó que se trabajará con una base de datos que consta de 9.272.800 (Cálculo de 13.400 registros multiplicados por 692 características) puntos de información para entrenamiento y prueba de los modelos, que permiten estudiar el problema con algoritmos de aprendizaje automático supervisado. Con esto, se da por cumplido el primero objetivo de la investigación relacionado con la identificación de los datos necesarios para el análisis de señales generadas a partir de imágenes espectroscópicas de resonancia magnética (MRSI) de los tejidos cerebrales comprometidos con GBM.

2.2 Aplicación de la resonancia magnética para el análisis de tumores cerebrales, conversión de las señales MRSI de Hertz a ppm e identificación de biomarcadores en tiempos de eco cortos

La resonancia magnética espectroscópica tiene como función la medida de los niveles de metabolitos en tejidos corporales, los cuales, pueden ser obtenidos a través de múltiples vóxeles (Delgado-Goñi et al., 2016). En nuestro caso, los pacientes a estudiar son ratones que hacen parte de la base de datos del Tumor Bank Repository del National Cancer Institute (Frederick, MD, USA). A estos individuos se les toman imágenes y éstas se clasifican dependiendo a qué grupo pertenecen de las etiquetas anteriormente descritas.

Cada MRSI extrae los espectros en los cuales se observan los niveles de metabolitos por vóxel. En nuestro caso de estudio, cada imagen es subdividida en una grilla de 10 x10 para un total de 100 vóxeles. Cada vóxel, podría representar una señal para el estudio del caso como parte de un procesamiento a partir de redes convolucionadas de una dimensión (Kiranyaz et al., 2021). Al mismo tiempo, un experto en radiología intenta clasificar cada uno de los vóxeles según criterio médico en las clases N, R, T y X, donde X, es una clase sin identificar y es lo que convierte el problema en semi-supervisado ya que, a partir del entrenamiento de los registros de las 3 primeras clases, podremos clasificar X en alguna de ellas. De tal forma que, como resultado, se puede conocer la clasificación de los tejidos en la MRSI inicial. Este proceso se puede observar en la figura 5; en donde se tiene el MRSI, posteriormente obtenemos dos elementos, los espectros por cada vóxel y la clasificación del expert y, posteriormente se espera generar una clasificación con los modelos definidos con el objetivo de delimitar las zonas sanas y aquellas comprometidas con tumor:

Figura 5 Flujo de la información



La importancia del proceso mostrado en la figura anterior radica en que se cuenta con la información espectroscópica de toda la imagen, pero solo con las etiquetas parciales suministradas por el experto. De tal forma, se necesita de un sistema automático para predecir en las zonas de frontera y que permita complementar la experiencia del profesional en aquellos casos donde le sea complejo identificar a qué tipo de tejido pertenece aquellos vóxeles que comparten áreas entre tumor y tejido sano y que, a simple vista no se tenga suficiente información, aprovechando de esta manera las características que pueda brindar la espectroscopía de cada zona y permite diferenciar entre clases de una manera acertada.

Luego de codificar el pipeline, usando técnicas para tomar la información original (sin manipular) desde distintos directorios en la memoria, se realizó el cargue de la información por directorios refiriéndose a cada grupo de ratones, siendo éstos A, B y C.

Después de cargar la información, se procedió a concatenar las tres bases de datos incluyendo el ID de cada ratón, las 692 características y las etiquetas N, R, T y X. Cada uno de los eventos proviene de la toma de información espectroscópica, que representa la actividad celular (metabólica) ocurriendo dentro del tumor.

Ahora bien, esta información espectroscópica es obtenida originalmente con datos entre un rango de -0.000786 y 0.144650 para la variable dependiente (Y) que corresponde a las etiquetas de cada uno de los registros, y de 0 a 691 para la variable independiente (X) que corresponde a las 692 características.

Ahora bien, esta información espectroscópica originalmente toma valores dentro de un rango de -0.000786 y 0.144650 en el eje (y) y, a su vez, va de 0 a 691 en el eje (x) describiendo así su comportamiento y dando información sobre 692 características.

Con el fin de alinear la investigación con la literatura actual (Govindaraju et al., 2000) y el estado del arte, fue necesario convertir los datos crudos de la información espectroscópica que

estaban en Hertz, a partes por millón - ppm. Para ello, se creó una función que toma las 692 características o puntos en este caso y se le asigna su correspondiente ppm.

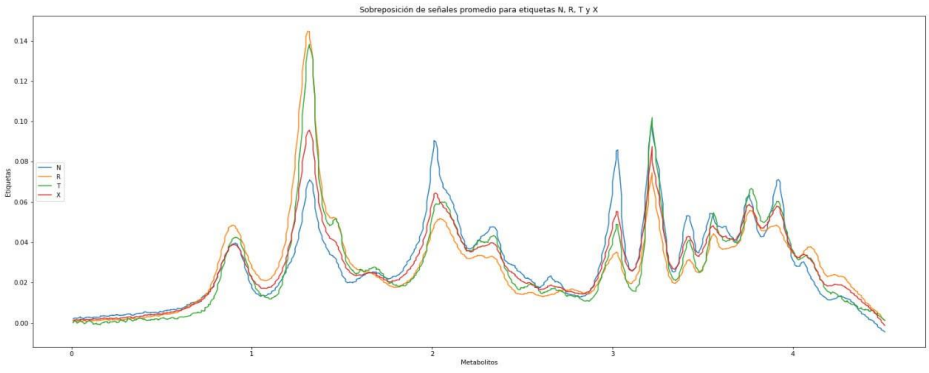
Esta conversión permite expresar los mismos valores de desplazamiento químico independientemente del espectrómetro utilizado en el estudio (KOPOT, n.d.) . Esta conversión es necesaria ya que evita que el desplazamiento químico sea estudiado en diferentes escalas dependiendo de la frecuencia del espectrómetro y su respectivo campo magnético. Convertir la frecuencia de la actividad metabólica de Hertz a ppm, estandarizará las unidades en que se estudia el mismo compuesto.

Para realizar esta conversión, se desarrolló un código para la conversión de las unidades descritas entre 0 y 691, a ppm entre un rango de 0 a 4.5. Cómo se observa en la figura 6, las partes por millón indican la posición del metabolito y, la altura del pico, su nivel de concentración en la zona. Aquí se puede observar que cada punto en el eje (x) del comportamiento de cada espectro será considerado como una característica de la clase, es decir, una variable. Con respecto al eje (y), representa la posición relativa de un metabolito a otro.

Adicionalmente, fue necesario aplicar el promedio para todos los eventos de cada etiqueta (N, R, T y X). Este promedio permite observar las 4 clases para los 13.400 datos del estudio y diferenciar gráficamente los patrones de comportamiento de los metabolitos en las MRSI.

Como resultado de este proceso, en la figura 6 se presenta el promedio de las señales por etiqueta en unidades por ppm.

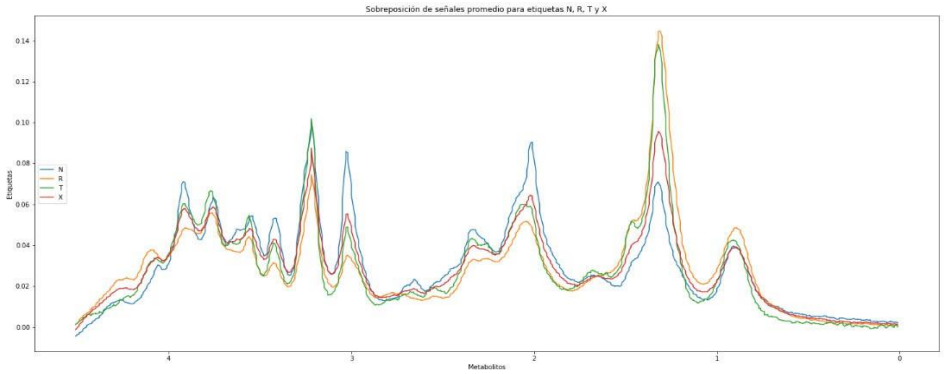
Figura 6 Señales promedio por etiqueta N (azul), R (naranja), T (verde) y X (rojo) en ppm



Para ser consecuentes con la literatura (Delgado-Goñi et al., 2016), se hizo una rotación de 180º en el eje vertical, obteniendo la figura 7. Uno de los primeros patrones que se observa es la tendencia de la etiqueta X, la cual pareciera mostrar un promedio entre las clases N, R y T.

Cabe recordar que esta etiqueta corresponde a la “No posible identificación del tejido” que hace el experto al no poder reconocer a cuál de las otras 3 clases pertenece la muestra en estudio.

Figura 7 *Señales promedio por etiqueta N (azul), R (naranja), T (verde) y X (rojo) en ppm, con rotación vertical de 180º*

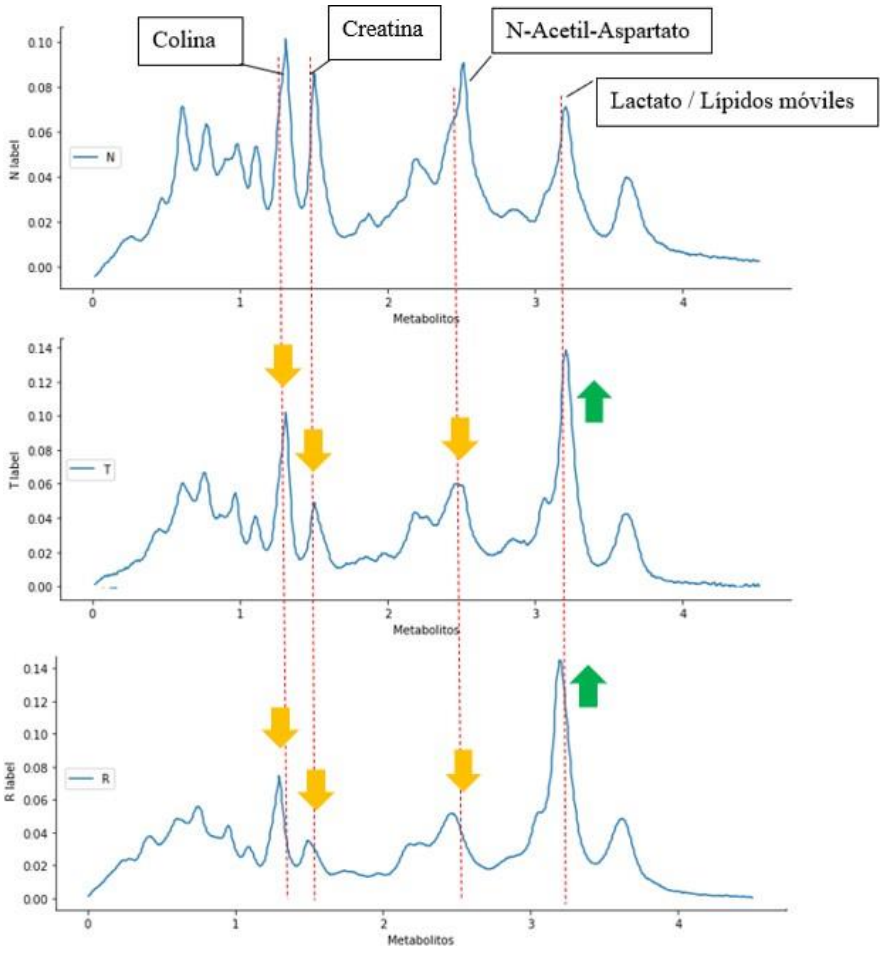


Una vez se determina la media de los registros para cada una de las etiquetas (N, R, T y X), podemos observar que las señales promedio derivadas de MRSI permiten discernir que hay algunos cambios fundamentales en los biomarcadores para incrementar o disminuir la altura relativa de los picos, por lo que resulta necesario observar en detalle qué tipo de biomarcadores son susceptibles a la presencia o no de tejido comprometido con GBM (Govindaraju et al., 2000).

De acuerdo con los estudios previos de GBM ya enunciados, se debe describir qué tipo de parámetros determinan la información extraída de los espectros, siendo uno de los más relevantes el tiempo de eco (TE time echo), el cual, está dividido en tiempo de eco largo y tiempo de eco corto. La principal diferencia entre estas dos categorías, consiste en que el primero provee espectros con una línea base menos distorsionada pero con una observación menor de metabolitos.

Por esta razón, continuamos nuestro estudio tomando como base el tiempo de eco corto (STE) y en la figura 8 se graficó el promedio de las observaciones por clase, comparando el comportamiento de los biomarcadores más relevantes (Colina-Cho, Creatina-Cr, N-Acetil-Aspartato-NAA y Lactato-Lac) en cada uno de los estados de los tejidos, tal como se había realizado en estudios previos sobre GBM (Govindaraju et al., 2000).

Figura 8 Comportamiento de los metabolitos en tiempo de eco corto



Fuente: Referencia estudio previo (X. P. Zhu et al., 2006)

En la figura 8, evidenciamos que los metabolitos más relevantes en el estudio de GBM. La concentración de Colina, Creatina y el N-Acetil-Aspartato disminuye cuando el tejido cambia de Normal a Tumoral. En cambio, el Lactato incrementa en la misma fase hacia el tumor (T). También existen cambios en los biomarcadores cuando cambia de Tumor en control (T) a Tumor en respuesta (R) ya que la presencia de la Colina y la Creatina desciende aún más cuando están en respuesta (R). El N-Acetil-Aspartato tiende a presentarse en similar concentración entre las clases que representan la existencia de tumor, de igual forma el Lactato mantiene su pico. Teniendo en cuenta esto, se podría inferir que los Biomarcadores Creatina y Colina nos permitirán diferenciar con mayor facilidad entre las clases (T) y (R) y, el N-Acetil-Aspartato y el Lactato, a clasificar entre las clases (N) y las clases involucradas con tumor.

2.3 Análisis exploratorio de datos

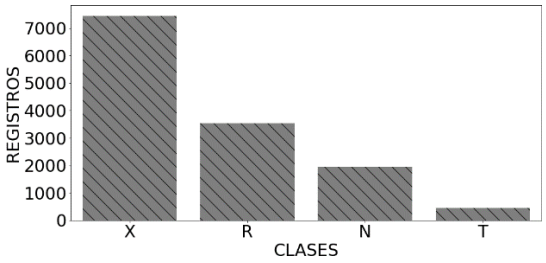
Este análisis se realizó con el fin de encontrar patrones preliminares de los datos, entender las dimensiones de la data, hacer análisis de tendencias centrales, conocer la distribución y dispersión de los datos.

A continuación, la tabla 2 y en la figura 9, se describen los porcentajes de cada una de las clases y sus posibles pesos.

Tabla 2 Tabulación cuantitativa de los datos por etiqueta N, R, T y X

ÍNDICE	CLASES	REGISTROS	PORCENTAJE (%)
0	X	7451	55.6
1	N	3544	26.45
2	R	1956	14.6
3	T	449	3.35

Figura 9 Tabulación cuantitativa de los datos por etiqueta N, R, T y X



Se observa en la tabla 2 y la figura 9, que la clase X posee la mayor cantidad de registros y tiende a desbalancear la distribución de los datos. Adicionalmente, se analiza que en puntos ppm particulares (*¨Peaks¨*), el promedio de los registros de la clase X siempre estará en algún punto entre el promedio de los registros de las clases R, N y T. Esto se debe a que la clase X eventualmente se podrá identificar como algunas de las otras tres clases. Es decir que en la clase X habrá comportamientos de los metabolitos como aquellos de las demás clases.

Teniendo en cuenta los objetivos del proyecto, se procede a concluir inicialmente que la base de datos es semi-supervisada ya que solo tres clases están completamente etiquetadas. La cuarta clase (X) no posee una etiqueta clara por lo que se elimina del conjunto de datos antes de avanzar con la minería de estos (aprendizaje automático supervisado).

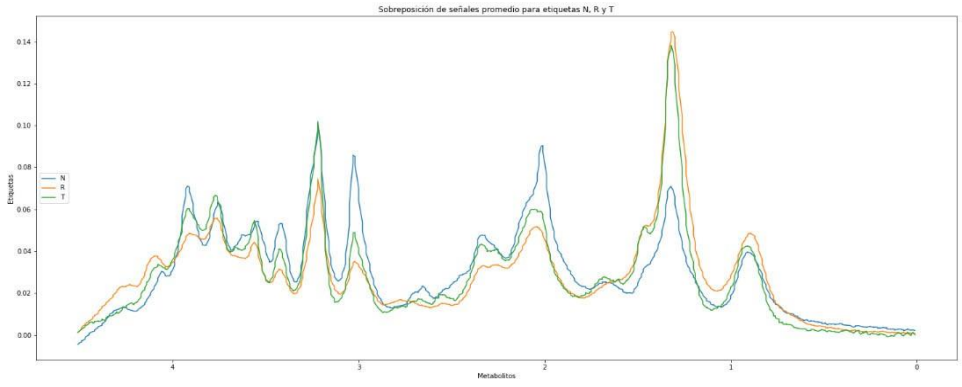
En este orden de ideas, en las bases de entrenamiento y testeo se eliminó la etiqueta con la clase X para la construcción del modelo pues esta no permitía reconocer patrones de los demás tejidos en estudio. Sin embargo, más adelante se tendrá en cuenta durante el desarrollo del trabajo. En el siguiente capítulo, procederemos a utilizar el mejor modelo para etiquetar la clase X y comparar estos resultados con los obtenidos en estudios previos de GBM referidos ya en este documento (1.1 y 1.3).

En la figura 10, se presentan las señales promedio de las tres etiquetas con las cuales se continuará el análisis: N, R y T. Aquí podemos observar picos que resultan claves para diferenciar las tres clases:

- La altura del lactato destaca a simple vista en las clases comprometidas con GBM (R y T).
- La creatina es superior en las zonas normales.

Esto nos permite reconocer que la espectroscopía podría brindar información valiosa para determinar aquellas características (que se traducen en información metabólica) que lleven a una clasificación acertada de las zonas a analizar y que a la vista podría ser un poco más complejo de etiquetar correctamente.

Figura 10 Señales promedio para etiquetas N (azul), R (naranja), T (verde) en ppm con rotación vertical de 180º



Dentro de la exploración de los datos observamos que las señales constan de 692 características, por lo que se opta por buscar una metodología que permita visualizar bases de

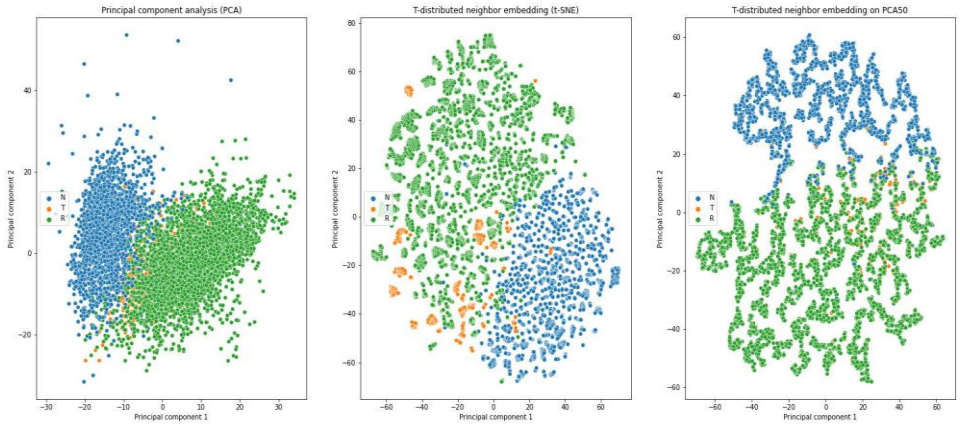
datos con reducción de dimensionalidad a dos dimensiones (plano XY) puesto que se podría interpretar de mejor manera los resultados teniendo en cuenta el tipo de problema que enfrentamos con el fin de encontrar inicialmente fronteras claras que permitan la clasificación de las clases. Para esto recurrimos a tres técnicas:

a. Técnica matemática de análisis de componentes principales (PCA) que permite explorar posibles correlaciones entre las variables a partir de la reducción de las dimensiones tratando de mantener la mayoría de la información de la base de datos original (Patel, 2019). Para ello, se utilizó la librería Scikit learn para reducir la dimensionalidad a 2 componentes principales y parametrizó el hecho de que la columna de las etiquetas se utilice para la codificación del color de las clases. El resultado se puede observar en la parte izquierda de la figura 11. Se evidencia que hay una clara diferencia entre los tejidos normales (N) y los tumorales, pero para estos últimos, no existe una división clara entre aquellos tumores en respuesta (R) o en control (T).

b. Técnica probabilística de implante de vecinos estocásticos t-Diseminados (t-SNE), la cual permite la visualización de base de datos con reducción de dimensionalidad, ya que recurre a localizar y representar gráficamente la distribución de la información multidimensional a un espacio de dimensiones más bajas y se concentra en identificar diseños de los datos en ramificaciones que se enfocan en la cercanía de la información (Boytsov et al., 2017). El resultado del t-SNE, que en la figura 11 corresponde al recuadro del medio, converge a la misma apreciación del análisis de componentes principales (PCA) toda vez que deja ver de mejor forma que las ramificaciones tumorales están mezcladas para las clases R y T, y por ende, no existe un borde que permita identificar claramente estas dos clases entre sí.

c. Las dificultades de visualización de las dos técnicas anteriores, nos llevó a continuar con la metodología y generar una tercera visualización aplicando la mezcla de las dos anteriores. Se practicó un t-SNE adicional, pero tomando como variables a visualizar, el resultado de las dos componentes del PCA. Esto permitió identificar aún con mayor certeza que las clases R y T se identifican mezcladas entre sí, tal como puede observarse en el tercer recuadro de la figura 11. De esta manera se puede concluir que, las características de ambas clases presentan elementos en común que al realizar una reducción de dimensionalidad podría dificultar la clasificación y brindar información poco útil al algoritmo para diferenciar los componentes de cada etiqueta.

Figura 11 Visualización de base de datos con reducción de dimensionalidad para etiquetas N (azul), R (verde) y T (naranja)



En consecuencia, se decide que la metodología para abordar el modelo no podría ser un problema de clasificación multiclase para 3 etiquetas (N vs R vs T) sino más bien, formular el problema a través de dos modelos binarios, por medio de la clasificación inicial de N vs RT, es decir, identificar si la zona se encuentra sana o está comprometida con tumor. Enseguida, los registros que sean clasificados como zona tumorar serán evaluados un segundo modelo binario que identificará si dicha zona es R o T. De esta forma, nuestra hipótesis sería que es menos complejo para los modelos identificar las clases de un problema binario y poder identificar con major precision las clases R y T concentrándose en encontrar las diferencias específicas entre ambas clases.

2.4 Selección de variables

Se efectuó un proceso de selección de variables con el ánimo de minimizar la función de costos por medio de un modelo lineal. A medida que la cantidad de datos incrementa, la calidad de los datos necesarios para el procesamiento mediante la minería de datos, el reconocimiento de patrones, el procesamiento de imágenes y otros algoritmos de aprendizaje automático disminuyen gradualmente. Bellman llama a este escenario "La maldición de la dimensionalidad" (Venkatesh & Anuradha, 2019)

El alto número de característica en una base de datos podría implicar la presencia de datos irrelevantes, redundantes y generadores de ruido, provocando problemas como sobreajuste y reduciendo la capacidad del algoritmo para realizar predicciones acertadas. Con el objetivo de prevenir los problemas mencionados, se suele optar por utilizar métodos de selección de características y extracción de características, como enfoques principales (Venkatesh & Anuradha, 2019).

En nuestro caso, estudiaremos el método Correlational based Feature Selection (CFS), que permite seleccionar aquellas características útiles para dar solución a la problemática. Este método funciona aplicando una heurística en donde se busca que el subconjunto de variables tenga una alta correlación con la variable objetivo y, no estén correlacionadas unas con otras. Para un grupo de variables discretas podría utilizarse la incertidumbre simétrica y si se tienen variables continuas se podría evaluar con la correlación de Pearson (Gopika & kowshalaya M.E., 2018)

Al realizar este proceso para los dos modelos propuestos, se obtuvo una reducción de dimensionalidad drástica, distribuida de la siguiente manera de acuerdo con el objetivo de clasificación de cada modelo:

- Modelo para clasificar clases con tumor (RT) y normal (N): [496, 484, 497, 493, 383, 494, 495, 384].
- Modelo para clasificar clases tumor con tratamiento (R) y sin tratamiento (T): [580, 594, 690, 664, 667, 687, 666, 671, 649, 689].

A pesar de las bondades que presenta este método y, en general el proceso de reducción de dimensionalidad; los modelos estructurados utilizando las variables seleccionadas no muestran mejoría en términos de sobreajuste y sacrifican el desempeño del algoritmo para distinguir entre clases. Esta situación se evidencia en especial a la hora de distinguir entre las clases de tumor con tratamiento (R) y sin tratamiento (T), puesto que son bastantes similares y conservar todas las características podría ser una necesidad para la capacidad de acertar del modelo propuesto.

3 Modelos de predicción sobre la clasificación de tejidos cerebrales comprometidos con GBM

En este capítulo, se da cumplimiento al tercer objetivo de esta investigación: diseñar modelos de predicción sobre la clasificación de tejidos cerebrales comprometidos con GBM, con métricas de desempeño confiables y reproducibles. Como se indicó en la metodología, para esta investigación se trabajaron los siguientes modelos de ML y DL:

- Métodos lineales: regresión logística y máquinas de soporte vectorial (SVM)
- Métodos no lineales: Random Forest y XGBOOST.
- Método de redes neuronales convolucionadas, y redes neuronales convolucionadas de una dimensión con el objetivo de aprovechar la forma de la señal.

Se procedió a utilizar una combinación de algoritmos que nos ofrezcan características distintas, permitiendo dar respuesta al problema desde distintos ángulos. En principio el algoritmo de regresión logística, que es uno de los más utilizados y a su vez nos brinda un acercamiento a las redes convolucionadas. Posteriormente, un SVM que introduce el concepto de hiperplano, buscando la mejor separación entre clases. Continuamos con dos algoritmos robustos basados en árboles de decisión, Random Forest y XG-Boost, en donde este último combina técnicas de ensamble y podría potenciar las capacidades de los árboles de decisión. Y, por último, la aplicación de redes convolucionadas de una dimensión que, se espera, permita clasificar con una mayor eficiencia que los modelos anteriormente descritos.

Se busca durante el desarrollo del trabajo poder estructurar una comparativa entre modelos de ML y DL que propicie un ejercicio de evaluación, determinando el mejor modelo dentro de los escogidos por su desempeño.

Para objetivos de mejorar la predicción ofrecida por los modelos de ML tradicionales, se tuvieron en cuenta conceptos importantes como *Cut-off point*, optimización de hiperparámetros y técnicas de balanceo de muestras. Con respecto a los modelos desarrollados usando redes neuronales convolucionadas de una dimensión, se considerarán las técnicas de balanceo de muestra y *early-stopping*.

Un parámetro importante dentro de los problemas de clasificación es el *Threshold* o *Cut-off point* de un modelo con el objetivo de mejorar la efectividad del diagnóstico. Esto representa la probabilidad de que una predicción sea cierta. Por defecto, este se encuentra en 0.5, lo cual puede variar según la necesidad que se plantea resolver. Las métricas de sensibilidad y especificidad dependen de escoger adecuadamente un punto de corte óptimo; por este motivo es necesario ser estimado por algunos criterios de optimización. De estos se destaca el índice de Youden que ha sido bastante utilizado en la práctica (Yin & Tian, 2014).

La sensibilidad se define como la capacidad de un modelo para identificar los casos que presentan la enfermedad correctamente. La especificidad indica la capacidad del modelo para identificar los casos sanos correctamente (Berrar, 2019). En el caso de nuestra investigación se busca balancear la relación entre sensibilidad y especificidad. Para esto utilizamos la métrica del *Yoden Index.*

El *Youden Index* indicará un mejor rendimiento en un determinado punto de corte si la sensibilidad y la especificidad fuesen igual de importantes o deseables dentro del modelo de clasificación (Bantis et al., 2014). De este modo, se espera que este índice sea lo más grande posible, es decir, que la diferencia entre sensibilidad y especificidad sea lo más pequeña, dando como resultado, una menor cantidad de falsos negativos y falsos positivos como resultado de la clasificación. Ergo, el objetivo principal de este ejercicio fue encontrar un valor que permitiera balancear las métricas mencionadas anteriormente. Se procederá a aprovechar las bondades de esta métrica para estimar el punto de corte con el que se deberá realizar la evaluación de los modelos de ML y, de esta manera asegurar que la clasificación de clases sea lo más adecuada posible al establecer la probabilidad con la que será identificada la clase RT (presencia de tumor) en el caso del primer modelo binario y, la clase R en el caso del segundo modelo.

La optimización de hiperparámetros resulta una tarea importante a la hora de estructurar modelos de predicción y clasificación. Encontrar la fórmula adecuada podría mejorar significativamente el desempeño del modelo. Los tres algoritmos más populares para lograr este cometido son: GridSearchC, RandomSearch y Bayesian Optimization. Para el primero es necesario establecer los hiperparámetros a optimizar junto con posibles valores, este algoritmo hará una búsqueda exhaustiva pasando por cada combinación posible de los valores que se destinaron. Esto lo hace bastante costoso en términos de esfuerzo computacional y tiempo. En el segundo, se establece una distribución estadística para cada parámetro y a partir de esto se muestrean los posibles valores. Aquí se puede fijar el número de modelos que se desean probar, lo que hará que utilice menos recursos. El tercero, es un algoritmo de optimización secuencial basado en modelos, es decir, que se apoya de la iteración anterior para definir los valores de los hiperparámetros de los modelos siguientes. Esto se repite hasta que converge en el modelo óptimo (Gupta, 2020).

Adicionalmente, para todos los modelos de ML y DL que se presentan a continuación, debe tenerse en cuenta que en el proceso del desarrollo del modelo óptimo nos encontramos con el problema de que la data no se encontraba balanceada, pues la clase R posee una cantidad mayor de observaciones que las clases N y T.

En este sentido y, con el objetivo de mejorar el resultado de la clasificación de los modelos, se propuso balancear el dataset. Existen dos procedimientos principales para lograr esta meta, sobremuestreo y submuestreo.

- El sobremuestreo busca incrementar la cantidad de muestras de la clase minoritaria. En este tipo de *data augmentation*, la técnica Synthetic Minority Oversampling Technique or SMOTE, es la más comúnmente utilizada. Esto permite que el modelo tenga la

capacidad de generalizar mejor y crear regiones de decisión más grandes, disminuyendo la posibilidad de que el modelo se sobreajuste a la clase con mayor número de registros. Con esta metodología se asegura la no pérdida de información, así como el mejoramiento del sobreajuste (T. Zhu et al., 2017). Sin embargo, para nuestra investigación se considera que podría introducir ruido al modelo.

- El submuestreo busca disminuir la cantidad de registros de la clase mayoritaria hasta obtener el mismo número de registros de la clase minoritaria. De igual manera como el anterior, se evita que exista un sobreajuste a la clase mayoritaria. Adicionalmente reduce el almacenamiento, implicando menor tiempo de ejecución. Sin embargo, la pérdida de información de la clase mayoritaria podría dar como resultado que la muestra elegida aleatoriamente no represente fielmente las características de la clase, minimizando la capacidad del modelo para identificar y clasificar información nueva (Mohammed et al., 2020).

Considerando esto, se descartaron las técnicas de submuestreo sobre la información, puesto que se reduciría considerablemente la cantidad de observaciones que utilizará el modelo para entrenarse y esto podría llevar a perder información relevante para clasificar con mayor precisión. De este modo, se optó por recurrir a SMOTE que incrementará la muestra de la clase minoritaria, en este caso N y T.

Por último, se referenciará la técnica de *early-stopping*, que es una función que permite que el proceso de entrenamiento del modelo de DL se detenga cuando se aprecie que la métrica que se haya escogido para este fin no mejore dentro del número de épocas determinado. A medida que los datos de entrenamiento son pasados por el algoritmo de aprendizaje en un determinado número de repeticiones se espera que este mejore, sin embargo, esto puede no ocurrir (Prechelt, 2012). Declarando un *early-stopping* indicando el momento en que debería detenerse, implicaría obtener el mejor modelo posible al evitar el sobreajuste y disminuir tiempos de ejecución.

Adicionalmente, es importante destacar que se intentará resolver la presencia del sobreajuste en los modelos de ML dividiendo los datos entre entrenamiento y testeo, de manera que un ratón no pertenezca a ambos conjuntos de datos. Esto debido a que nace la premisa de que podría causar *data leakage* (se define como fuga de datos al brindar información al conjunto de testeo sobre el conjunto de entrenamiento). Otro método es definiendo hiperparámetros óptimos que beneficien el desempeño del modelo, así como el cálculo de un punto de corte óptimo. Con respecto a los modelos de DL, este objetivo se buscará por medio de la regularización.

3.1 Modelo de regresión logística para problemas de clasificación de tejidos multiclase

Para el caso de nuestra investigación, se desarrollaron dos modelos de clasificación haciendo uso del algoritmo de regresión logística. Se utilizó la información de los ratones de las clases N,

R y T, distribuida aleatoriamente en el conjunto de entrenamiento y testeo, de manera que un ratón no pudiese encontrarse en ambos conjuntos de datos. Para el primer modelo (N vs RT) se contó con 4.245 registros en entrenamiento y 1.704 registros en testeo. En este ejercicio RT será la clase positiva y N la clase negativa. Para el segundo modelo (R vs T) se tuvo en cuenta solamente la información de las clases R y T. Por tal motivo se contó con 2.666 registros para entrenamiento y 1.327 para testeo. R será la clase positiva y T, la clase negativa.

Adicionalmente hay que aseverar que son 692 características con las que se trabajará finalmente, puesto que los intentos de reducción de dimensionalidad no mostraban el desempeño esperado y por este motivo se procedió a utilizar la información completa.

Posteriormente se procede a utilizar SMOTE para balancear el conjunto de datos y proceder el entrenamiento del modelo, definiendo los hiperparámetros óptimos que permitieran obtener modelos más robustos y eficientes.

Para cada uno de los modelos desarrollados, se tuvieron en cuenta los siguientes hiperparámetros, puesto se consideran los más pertinentes y la razón se explica a continuación (Pedregosa et al., 2011):

- Solver: se define como el algoritmo que se utilizará el problema de optimización.
- Max iter: significa la cantidad máxima de iteraciones que se permite para que el solver pueda converger.
- Penalty: se encuentra asociada con la regularización del modelo. Es una forma de evitar el sobreajuste al reducir los parámetros y simplificar el modelo. Funciona al agregar penalizaciones a los modelos con mayor complejidad y se escoge el modelo que tenga la puntuación más baja de sobreajuste. Depende del solver que se escoja en el modelo puesto que no todos soportan cada solver que existe. (Pekhimenko, 2006)
- L1 va a limitar el tamaño de los coeficientes del modelo. Asigna coeficientes cero a algunas variables con menor contribución para el modelo. Se mantienen las variables significativas para el modelo.
- L2 va a asignar un valor cercano a cero a aquellas variables que menos contribuyen al modelo, lo que permite conservar la participación de todas las variables.
- Elasticnet es la combinación entre la regularización L1 y L2. Se encarga de establecer algunos coeficientes cercanos a cero y otros iguales a cero.
- L1_ratio: al encontrar como óptimo el solver saga y penalty elasticnet, este parámetro podría ofrecer una mejor optimización.
- C: este parámetro va a precisar la fuerza de la regularización en el modelo. El valor de C y la regularización se encuentran negativamente correlacionados. Entre más pequeño sea C, más fuerte será la regularización.

Se plantea un primer modelo que permitiera detectar la presencia de tumor en una zona determinada. Para efectos de esto, se utilizó un algoritmo con las siguientes características:

- Solver: En este caso se utilizó ***saga***, que presenta grandes ventajas para bases de datos robustas. Es una variante del algoritmo ***sag,*** Stands for Stochastic Average Gradient Descent. Saga permite utilizar la regularización L1.
- Max iter: En este caso, se requirió un número máximo de 20.000 iteraciones.
- Penalty: La regularización óptima es *elasticnet.*
- L1_ratio: El mejor valor encontrado fue 0.5, lo que significa que la penalidad podría ser una combinación entre l1 y l2 y adicional, define el peso que tendrá l1 al efectuarse esto.
- C: 545.55947

Posteriormente, se calculó el *Threshold* óptimo que tendría este modelo que es **0.7625**, lo que significa que las predicciones que posean una probabilidad superior o igual al 0.7625 serán clasificados en la clase tumoral (RT) y, en caso contrario, serán determinados como la clase normal (N).

Luego, se estructuró un modelo que tuviese como objetivo determinar si un tumor había recibido tratamiento (R) o no (T). Para llevar a cabo esto, se tuvieron en cuenta los siguientes parámetros:

- Solver: sag.
- Max iter: 10.000.
- Penalty: La regularización óptima es *l2*.
- C: 1438.4

El Threshold óptimo en este caso fue de **0.985**, lo que significa que las predicciones que posean una probabilidad superior o igual al 0.985, serán clasificados como un tumor con tratamiento (R). En caso contrario, será un tumor sin tratamiento (T).

La elección de los hiperparámetros óptimos se realizó con el apoyo de GridSearchCV, que es una técnica de validación cruzada, que permite escoger la mejor combinación de valores para los hiperparámetros que sean considerados para evaluación. Se procedió a analizar los hiperparámetros más relevantes dentro de cada algoritmo y de este modo, ajustar valores sensibles dentro del abanico de posibilidades que tuviese esta metodología para encontrar los valores que mejor se ajustaran al problema en cuestión. Su búsqueda es exhaustiva y representa un costo computacional importante, sin embargo, se evidencian óptimos resultados en su utilización, en comparación con otras técnicas como RandomizedSearchCV que hace una búsqueda aleatoria.

Se evidencia que ambos modelos siguen presentando sobreajuste, sobre todo en el segundo modelo. Esta situación mejora cuando se utiliza el *Threshold* óptimo y la optimización de hiperparámetros, sin embargo, no desaparece. Otra situación que se vio beneficiada por lo anteriormente mencionado fue la sensibilidad y especificidad del modelo.

Adicionalmente se debe destacar que fue necesario la utilización de técnicas de *oversampling* para garantizar que los modelos no tuvieran sobreajuste sobre la clase con mayor número de registros.

3.1.1 Resultados

Tal como se muestra en las siguientes tablas, la regresión logística presenta un buen desempeño en el trabajo de clasificación, especialmente para el modelo 1, en donde se evidencia una cantidad más baja de falsos positivos y negativos. En la tabla 3 tenemos la matriz de confusión del modelo 1, en donde se puede concluir que el modelo realiza un buen trabajo de clasificación. En la tabla 4, en donde se encuentra la matriz de confusión del modelo 2, se infiere que al algoritmo le cuesta un poco más la diferenciación entre las clases R y T.

Tabla 3 Matriz de confusión modelo 1 regresión logística

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1276*	*51*
NEGATIVO ACTUAL	*8*	*369*

Tabla 4 Matriz de confusión modelo 2 regresión logística

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1050*	*83*
NEGATIVO ACTUAL	*29*	*165*

3.2 Modelo de máquinas de soporte vectorial para problemas de clasificación de tejidos multiclase – SVM

Para aplicar SVM a nuestra investigación se utilizó la información de los ratones de las clases N, R y T, distribuida aleatoriamente en el conjunto de entrenamiento y testeo, de manera que un ratón no pudiese encontrarse en ambos conjuntos de datos. Para el primer modelo (N vs RT) se contó con 4.245 registros en entrenamiento y 1.704 registros en testeo. En este ejercicio RT será la clase positiva y N la clase negativa. Para el segundo modelo (R vs T) se tuvo en cuenta solamente la información de las clases R y T. Por tal motivo se contó con 2.666 registros para entrenamiento y 1.327 para testeo. R será la clase positiva y T, la clase negativa.

Adicionalmente hay que aseverar que son 692 características con las que se trabajará finalmente, puesto que los intentos de reducción de dimensionalidad no mostraban el desempeño esperado y por este motivo se procedió a utilizar la información completa.

Posteriormente se procede a utilizar SMOTE para balancear el conjunto de datos y proceder el entrenamiento del modelo, definiendo los hiperparámetros óptimos que permitieran obtener modelos más robustos y eficientes.

Se desarrollaron dos modelos con las siguientes características (*Support Vector Machine (SVM) Hyperparameter Tuning In Python*, 2022).

- Kernel: corresponde a la función matemática utilizada en la transformación de los datos. Este puede ser linear, poly, rbf, sigmoid, precomputed. Su función principal es separar correctamente las clases para generar el hiperplano que permitirá efectuar la clasificación.
- Gamma: es el coeficiente del kernel si este es rbf, poly o sigmoid. Define la influencia de un solo punto de entrenamiento, un alto valor de este parámetro indica que los puntos sobre el plano deben encontrarse más cercanos para ser considerados como un grupo y, un valor más bajo implica que el radio de distancia entre puntos puede ser mayor para ser considerados como una misma clase.
- C: Es el valor que define la penalidad sobre la clasificación errónea del modelo. Es inversamente proporcional a la regularización, entre más pequeño sea este valor, el clasificador buscará un hiperplano que maximice el margen de separación entre clases y se tendrá una tasa mayor de errores en la clasificación.
- Degree: se ve involucrado cuando el kernel es poly. Establece la flexibilidad del clasificador para poder separar las clases del modelo.

El primer modelo busca clasificar las clases normal (N) y tumor (RT). Luego de su optimización presenta las siguientes características:

- Kernel: rbf
- Gamma: 0.5
- C: 10
- Threshold: 0.64

El segundo modelo, que clasifica las clases (R) y (T), se estructuró con los siguientes parámetros:

- Kernel: poly
- Gamma: 0.1
- Degree: 3
- C: 1000
- Threshold: 0.99

La elección de los hiperparámetros óptimos se realizó con el apoyo de GridSearchCV, que es una técnica de validación cruzada, que permite escoger la mejor combinación de valores para

los hiperparámetros que sean considerados para evaluación. Se procedió a analizar los hiperparámetros más relevantes dentro de cada algoritmo y de este modo, ajustar valores sensibles dentro del abanico de posibilidades que tuviese esta metodología para encontrar los valores que mejor se ajustaran al problema en cuestión. Su búsqueda es exhaustiva y representa un costo computacional importante, sin embargo, se evidencian óptimos resultados en su utilización, en comparación con otras técnicas como RandomizedSearchCV que hace una búsqueda aleatoria.

Se evidencia que el kernel óptimo de ambos modelos es distinto, esto puede deberse a la naturaleza de las características comunes y distintas entre las clases involucradas. Adicionalmente, cabe notar que el *threshold* óptimo del segundo modelo se encuentra bastante cercano a uno.

3.2.1 Resultados

Se observa que el modelo realiza una mejor predicción toda vez que presenta menos falsos negativos y positivos que el modelo de regresión logística tanto para el modelo 1 como para el modelo 2, lo anterior se puede observar en las matrices de confusión de las tablas 5 y 6, en donde se aprecia una mejora en el desempeño haciendo uso de SVM. Al igual que el algoritmo anterior, se evidencia una alta cantidad de falsos positivos y falsos negativos, tal como se expone en las siguientes tablas.

Tabla 5 *Matriz de confusión modelo 1 SVM*

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1296*	*31*
NEGATIVO ACTUAL	*8*	*369*

Tabla 6 *Matriz de confusión modelo 2 SVM*

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1052*	*81*
NEGATIVO ACTUAL	*22*	*172*

3.3 Modelo de random forest para problemas de clasificación de tejidos multiclase

Para hacer uso del algoritmo de Random Forest, se utilizó la información de los ratones de las clases N, R y T, distribuida aleatoriamente en el conjunto de entrenamiento y testeo, de manera que un ratón no pudiese encontrarse en ambos conjuntos de datos. Para el primer modelo (N

vs RT) se contó con 4.245 registros en entrenamiento y 1.704 registros en testeo. En este ejercicio RT será la clase positiva y N la clase negativa. Para el segundo modelo (R vs T) se tuvo en cuenta solamente la información de las clases R y T. Por tal motivo se contó con 2.666 registros para entrenamiento y 1.327 para testeo. R será la clase positiva y T, la clase negativa.

Adicionalmente hay que aseverar que son 692 características con las que se trabajará finalmente, puesto que los intentos de reducción de dimensionalidad no mostraban el desempeño esperado y por este motivo se procedió a utilizar la información completa.

Posteriormente se procede a utilizar SMOTE para balancear el conjunto de datos y proceder el entrenamiento del modelo, definiendo los hiperparámetros óptimos que permitieran obtener modelos más robustos y eficientes.

En esta investigación, se tuvo en cuenta que para construir los modelos de random forest es esencial contar con las siguientes características que garanticen el mejor desempeño de los resultados (Koehrsen, 2018)

- N_estimators: corresponde al número de árboles que utilizará el algoritmo.
- Criterion: define la función para medir la calidad de la división de los datos en un árbol de decisión.
- Max_depth: indica la máxima profundidad del árbol.
- Max_features: representa el número máximo de características consideradas al dividir un nodo.
- Bootstrap: indica si se utilizará una muestra de datos diferente para la creación de los modelos dentro del algoritmo de random forest.
- Min_samples_split: se refiere al número mínimo de muestras que estarán en un nodo hoja.
- Min_sample_leaf: establece el número mínimo de muestras requeridas en cada nodo del árbol de decisión.

En este sentido, para el primer modelo desarrollado para determinar la clasificación de zona con tumor (RT) y zona sin presencia de tumor (N) presenta la siguiente estructura:

- N_estimators: 1000.
- Criterion: entropy
- Max_depth: 40
- Max_features: auto
- Bootstrap: False
- Min_samples_split: 2
- Min_sample_leaf: 1
- Threshold: 0.559

El segundo modelo se construyó para clasificar una zona tumoral que ha recibido tratamiento (R) de la que no (T), siguiendo los siguientes parámetros:

- N_estimators: 100.
- Criterion: gini
- Max_depth: 70
- Max_features: stqr
- Bootstrap: False
- Min_samples_split: 10
- Min_sample_leaf: 2
- Threshold: 0.92

La elección de los hiperparámetros óptimos se realizó con el apoyo de GridSearchCV, que es una técnica de validación cruzada, que permite escoger la mejor combinación de valores para los hiperparámetros que sean considerados para evaluación. Se procedió a analizar los hiperparámetros más relevantes dentro de cada algoritmo y de este modo, ajustar valores sensibles dentro del abanico de posibilidades que tuviese esta metodología para encontrar los valores que mejor se ajustaran al problema en cuestión. Su búsqueda es exhaustiva y representa un costo computacional importante, sin embargo, se evidencian óptimos resultados en su utilización, en comparación con otras técnicas como RandomizedSearchCV que hace una búsqueda aleatoria.

Se evidencia la presencia de sobreajuste de los datos de entrenamiento en ambos modelos. La elección de un threshold óptimo en el primer modelo no dista en gran magnitud del valor 0.5 que es generalmente utilizado. En cuanto al segundo modelo, aunque el threshold óptimo es inferior al que se evidencia en otros modelos, sigue siendo cercano a 1.

3.3.1 Resultados

El algoritmo de random forest demuestra especial destreza para identificar los casos normales y con tumor, como se observa en la tabla 7. Esto no se evidencia en el modelo para clasificar una zona con tratamiento vs sin tratamiento, en donde demuestra ser el modelo con menor desempeño entre los desarrollados durante el trabajo, al dar como resultado 190 casos cuya predicción era la clase (T) y la etiqueta correcta era (R), como se muestra en la tabla 8.

Tabla 7 Matriz de confusión modelo 1 Radom Forest

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1296*	*31*
NEGATIVO ACTUAL	*4*	*373*

Tabla 8 Matriz de confusión modelo 2 Radom Forest

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*943*	*190*
NEGATIVO ACTUAL	*16*	*178*

3.4 Modelo de ensamble utilizando múltiples algoritmos de aprendizaje para problemas de clasificación de tejidos multiclase

Se dispone a utilizar el algoritmo de Random Forest en donde se utilizó la información de los ratones de las clases N, R y T, distribuida aleatoriamente en el conjunto de entrenamiento y testeo, de manera que un ratón no pudiese encontrarse en ambos conjuntos de datos. Para el primer modelo (N vs RT) se contó con 4.245 registros en entrenamiento y 1.704 registros en testeo. En este ejercicio RT será la clase positiva y N la clase negativa. Para el segundo modelo (R vs T) se tuvo en cuenta solamente la información de las clases R y T. Por tal motivo se contó con 2.666 registros para entrenamiento y 1.327 para testeo. R será la clase positiva y T, la clase negativa.

Adicionalmente hay que aseverar que son 692 características con las que se trabajará finalmente, puesto que los intentos de reducción de dimensionalidad no mostraban el desempeño esperado y por este motivo se procedió a utilizar la información completa.

Posteriormente se procede a utilizar SMOTE para balancear el conjunto de datos y proceder el entrenamiento del modelo, definiendo los hiperparámetros óptimos que permitieran obtener modelos más robustos y eficientes.

En este trabajo de investigación, los modelos desarrollados presentan las siguientes características con el fin de mejorar su desempeño en la clasificación:

- Booster: define el tipo de modelo que se utilizará en cada iteración. Se tienen tres opciones: gbtree (modelo basado en árboles de decisión), gblinear (modelos lineales) y dart (modelo basado en árboles de decisión que involucra dropout).
- Eta: mejor conocido como Learning rate. Es un factor de ponderación para los nuevos modelos resultantes de cada ponderación y así evitar que los datos se sobreajusten a los datos de entrenamiento.
- Min_child_weight: se refiere a la suma mínima de pesos de todas las observaciones para hacer una participación adicional en un nodo del árbol. Un valor muy alto podría evitar que el modelo aprenda características muy específicas del dataset de entrenamiento, disminuyendo el sobreajuste.
- Max_depth: define la profundidad máxima del árbol. Un valor alto podría evitar el sobreajuste de los datos a la data de entrenamiento.

- Subsample: es un valor entre 0 y 1 que representa la fracción de información aleatoria que se requiere para entrenar cada árbol iterado. Un valor bajo haría que el modelo fuese más conservador.
- Gamma: indica el valor mínimo de pérdida que se requiere para que se haga la siguiente división del nodo del árbol. Podría hacer que el modelo fuera más conservador.

El primer modelo propuesto para determinar la clasificación de zona con tumor (RT) y zona sin presencia de tumor (N) se constituye a partir de las siguientes características:

- Booster: gbtree
- Eta: 0.1
- Min_child_weight: 5
- Max_depth: 6.
- Subsample: 0.7
- Threshold: 0.63

El segundo modelo que se plantea como objetivo clasificar de una zona tumoral si efectivamente ha recibido tratamiento (R) o no (T), presenta los siguientes hiperparámetros:

- Booster: gbtree
- Eta: 0.15
- Min_child_weight: 1
- Max_depth: 7
- Subsample: 0.8
- Threshold: 0.992

En estos modelos se mantiene la tendencia a sobreajuste en los modelos previos. El Threshold óptimo del segundo modelo es cercano a 1. La elección de cada punto de corte óptimo implica una mejora tanto en el balanceo de los falsos positivos y negativos como en el sobreajuste en los datos de entrenamiento.

La elección de los hiperparámetros óptimos se realizó con el apoyo de GridSearchCV, que es una técnica de validación cruzada, que permite escoger la mejor combinación de valores para los hiperparámetros que sean considerados para evaluación. Se procedió a analizar los hiperparámetros más relevantes dentro de cada algoritmo y de este modo, ajustar valores sensibles dentro del abanico de posibilidades que tuviese esta metodología para encontrar los valores que mejor se ajustaran al problema en cuestión. Su búsqueda es exhaustiva y representa un costo computacional importante, sin embargo, se evidencian óptimos resultados en su utilización, en comparación con otras técnicas como RandomizedSearchCV que hace una búsqueda aleatoria.

3.4.1 Resultados

El primer modelo representa de una manera acertada las clases normal y con tumor. La cantidad de falsos positivos y falsos negativos no llegan a ser el 3,11% de las observaciones totales, sin embargo, no se aprecia una mejora sustancial con respecto a modelos anteriores, esto se obtiene con la información reflejada en la tabla 9. Con respecto al modelo 2, en la tabla 10 se aprecia que el número de falsos negativos es una suma importante del total de casos ydisminuye la confianza en la estructuración del modelo.

Tabla 9 Matriz de confusión modelo 1 XGBOOST

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1281*	*46*
NEGATIVO ACTUAL	*7*	*370*

Tabla 10 Matriz de confusión modelo 2 XGBOOST

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1008*	*125*
NEGATIVO ACTUAL	*26*	*168*

3.5 Modelo redes neuronales de dos dimensiones para problemas de clasificación de tejidos multiclase (2D CNN)

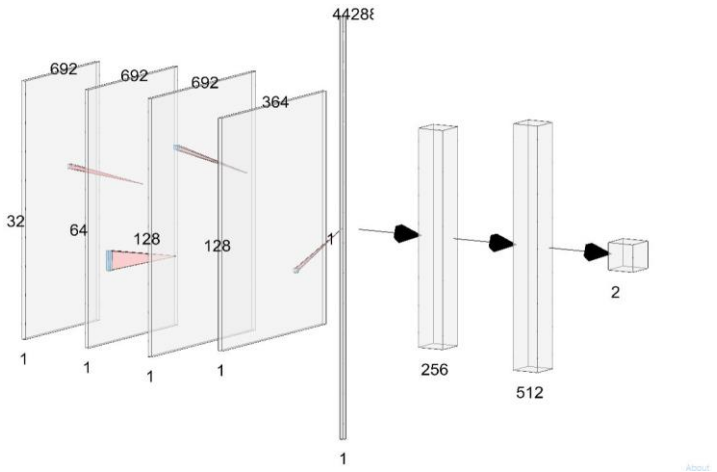
Se considera que, para nuestra investigación resulta pertinente este modelo, toda vez que se aplican redes convolucionadas a un problema tabular, originado en señales espectroscópicas.

Una de las alternativas planteadas durante el proceso de investigación fue tomar la gráfica resultante de colocar las señales espectroscópicas en un plano XY y aplicar redes convolucionadas de dos dimensiones (2D CNN) esperando el reconocimiento de patrones. La arquitectura utilizada para llevar esto a cabo (figura 12) se compone de varias capas convolucionadas, las cuales están conformadas por filtros que extraen patrones complejos innatos de las señales en estudio. A partir de esto, las señales se procesan como gráficas en el plano XY y se utiliza una arquitectura con capas de convolución con la siguiente descripción:

- Se normalizan los datos para las 692 variables, luego se aplica una capa densa de 16 neuronas, se aplica un dropout, una segunda capa densa, también de 16 neuronas, un segundo dropout, una tercera capa densa de 8 neuronas, otro dropout, y por último una capa densa de una neurona. Convirtiéndo la salida, en un problema binario para la aplicación de la función Sigmoidal.

Al aplicar la arquitectura (Hortúa, 2022) a los datos preprocesados se observa que el desbalance de las clases (N, R y T) no permite que el modelo converja a una solución del problema planteado. Para superar esto, se aplican técnicas de sobre muestreo para así balancear las categorías de clasificación que no cuentan con una representación equitativa en el dataset (Chawla et al., 2002).

Figura 12 Arquitectura para 2D CNN aplicada al caso de estudio



Con este nuevo procedimiento, se observa que las redes convolucionadas continúan sin converger a una solución factible ya que las métricas describen un sobre ajuste del modelo. A partir de esto, se procede a iterar la modificación de los pesos de las clases de acuerdo con unos porcentajes estipulados según se observa en la tabla 11:

Tabla 11 Comparación de métricas para arquitectura 2D CNN

Models	FalseNegative	Accuary Validation	LossTraining	LossValidation
Imbalance	0	0.6512	0.5704	0.5476
Bias Initializer	0	0.6512	0.6478	0.6477
SMOTE	11	0.9732	0.0747	0.0718
Weight 1.43-0.77	9	0.978	0.0947	0.0605
Weight 0.1-0.9	15	0.978	0.0836	0.0565
Weight 0.2-0.8	34	0.9535	0.667	0.1367
Weight 0.3-0.7	13	0.9819	0.0902	0.0572
Weight 0.4-0.6	15	0.9767	0.097	0.0652
Weight 0.5-0.5	12	0.9755	0.1013	0.0611
Weight 0.6-0.4	5	0.9742	0.0925	0.0687
Weight 0.7-0.3	4	0.9638	0.0934	0.0918
Weight 0.8-0.2	4	0.9574	0.1423	0.1048
Weight 0.9-0.1	13	0.9767	0.0975	0.0549

Se observa que aún con la aplicación de balanceo de las clases menos representadas del dataset y de la ponderación e iteración de los pesos de las clases, todos los modelos presentan

sobreajuste en las épocas iniciales del entrenamiento por lo que es complejo utilizar otras técnicas para tratar de evitar problemas. Así las cosas, se decide descartar completamente este modelo y no se comparará con otros en el proceso de evaluación.

3.5.1 Resultados

Como se muestra en la matriz de confusión del modelo 2D CNN (tabla 12), este procedimiento no permite predecir acertadamente a qué clase pertenece una nueva observación no antes vista, ya que, al calcular las métricas anteriores, observamos que no existen verdaderos positivos (TP) y falsos negativos (FN) lo que entendemos no es correcto según la literatura (Shultz et al., 2011).

Tabla 12 Matriz de confusión modelo 2D CNN

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*0*	*256*
NEGATIVO ACTUAL	*0*	*509*

3.6 Mapa de clasificación sobre puesto en los vóxeles para visualización de las clases de tejido cerebral a predecir

Teniendo en cuenta que las predicciones del modelo no fueron acertadas por sobre ajuste durante el entrenamiento y no se obtuvo un resultado ajustado a la realidad, no fue posible aplicar el mapa de clasificación inicialmente propuesto en la investigación.

3.7 Modelo de redes neuronales convolucionadas de una dimensión para problemas de clasificación de tejido cerebral multiclase

Los datos utilizados para ejecutar el entrenamiento de este algoritmo corresponden a la información de las clases N, R y T. Utilizando la librería Sklearn se procede a realizar la distribución aleatoria de los datos de entrenamiento y testeo. De esta manera, para el primer modelo, (N vs RT), se contó con 4.164 registros en el conjunto de entrenamiento y 1.785 registros de testeo, aproximadamente el 70% de información para el primero y 30% para el segundo. En este ejercicio RT será la clase positiva y N la negativa. En el caso del segundo modelo (R vs T), 2.795 en entrenamiento y 1197 en testeo, manteniendo la distribución anterior. R será la clase positiva y T la negativa.

Las redes convolucionadas de una dimensión siguen el mismo principio teórico de las 2D CNN al tratar de extraer características espaciales de la base de datos utilizando su Kernel. La gran diferencia radica en que para las 1D CNN, el Kernel se desliza en una dimensión mientras

que en las 2D CNN, el Kernel lo hace en dos dimensiones (Largo y ancho o XY) (Verma, 2019).

En este sentido, podemos inferir que, para nuestra investigación, el comportamiento de los metabolitos se les dará el tratamiento por medio del procesamiento de señales ya que no se limita a un conjunto específico de valores y puede variar sin necesidad de interrumpir la continuidad. Estos valores, se presentan como secuencias de puntos de información en un intervalo (en este caso, no se trata de un rango de tiempo) que nos permite realizar seguimiento a los cambios en la captura o respuesta de los metabolitos en el rango de estudio estipulado (vóxel). Inmediatamente, podemos reconocer que un kernel de 1D puede deslizarse sobre estas señales (serie de datos) ya que tiene propiedades espaciales, consideradas dentro del espectro.

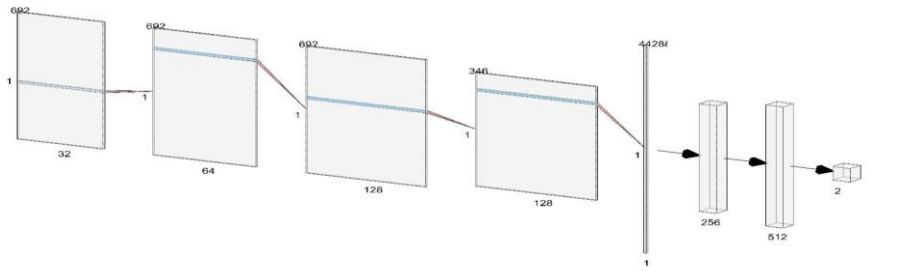
A continuación, aplicamos una arquitectura de 1D CNN (Premanand, 2021) a nuestros casos de estudio, haciendo las modificaciones de acuerdo con las dimensiones originales que servirán tanto para la entrada del modelo, como para las capas densas finales, en especial la de salida, para que se ajuste a arrojar el resultado para un problema binario por medio de dos neuronas. (Ver tabla 13 y Figura 13).

Tabla 13 Arquitectura de 1D CNN aplicada al caso de estudio

N°	Tipo	Observaciones
1	Conv 1D	N° filtros: 32, Tamaño filtro:3, Función de activación: ReLU, padding: same
2	Conv 1D	N° filtros: 64, Tamaño filtro:3, Función de activación: ReLU, padding: same
3	Conv 1D	N° filtros: 128, Tamaño filtro:3, Función de activación: ReLU, padding: same
4	MaxPooling	Tamaño región submuestreo: 3
5	Dropout	Tasa: 0.5
6	Flatten	
7	Densa 1	Unidades: 256, Funcion de activación: ReLU
8	Densa 2	Unidades: 512, Funcion de activación: ReLU
9	Densa 3	Unidades: 2, Funcion de activación: Softmax

Durante la elaboración del pipeline para esta modelo, observamos que la red convolucionada de una dimensión es sensible al desbalanceo de los datos, por lo que fue necesario aplicar técnicas de bajo y/o sobre dimensionamiento de las clases a los datos de entrenamiento, más no, a los de testeo. Este hallazgo, nos permitió producir un modelo con un menor costo computacional.

Figura 13 Arquitectura para 1D CNN aplicada al caso de estudio



Con este nuevo procedimiento, se observa que las redes convolucionadas de una dimensión si convergen a una solución factible ya que las métricas describen que no existe un sobre ajuste del modelo cuando se aplican técnicas como early-stopping.

3.7.1 Resultados

Teniendo en cuenta la matriz de confusión (tablas 14 y 15), se aprecia que tanto el modelo 1 y el modelo 2 demuestran una buena capacidad para clasificar nuevas observaciones de manera correcta. En ambos modelos se buscó por medio de la técnica de early-stopping y model checkpoints, encontrar y conservar el modelo cuyo valor de pérdida en los datos de validación sea el menor y no presente sobreajuste sobre los datos de entrenamiento.

Tabla 14 Matriz de confusión modelo 1 por 1D CNN

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1211*	*8*
NEGATIVO ACTUAL	*4*	*562*

Tabla 15 Matriz de confusión modelo 2 por 1D CNN

	PREDICCIÓN POSITIVA	PREDICCIÓN NEGATIVA
POSITIVO ACTUAL	*1051*	*2*
NEGATIVO ACTUAL	*1*	*144*

3.8 Mapa de clasificación sobrepuesto en las imágenes MRSI para visualización de las clases de tejido cerebral a predecir por medio de 1D CNN

Con el propósito de visualizar las zonas sanas (N) y afectadas por tumor (RT) de cada ratón se decidió ilustrar la predicción efectuada por el modelo de redes convolucionadas de una dimensión utilizando mapas de calor.

Inicialmente presenta un mapa de clasificación siguiendo las etiquetas de los expertos, en donde el color azul representa la case normal (N), el color naranja tumor sin tratamiento (T), el color verde tumor con tratamiento (R) y el color amarillo es una zona que no fue posible clasificar (X).

En un segundo momento se planteó lo siguiente: los datos de cada ratón son pasados a través del primer modelo desarrollado utilizando 1D CNN. El resultado arrojado es un mapa de la zona cerebral que clasifica cada vóxel (cuya información proviene del espectro) en la clase normal (N) cuyo color será azul, y la clase tumoral (R y T) cuyo color será naranja.

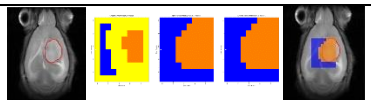
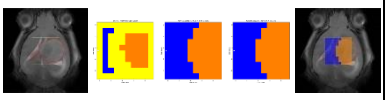
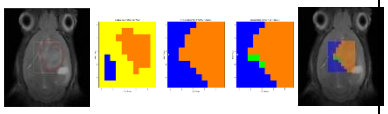
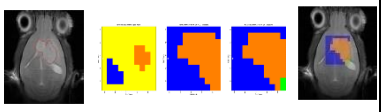
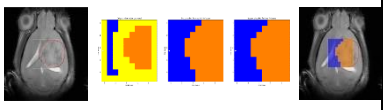
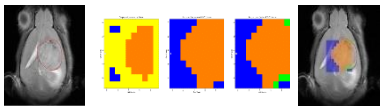
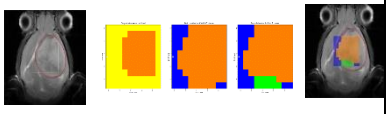
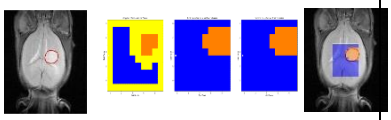
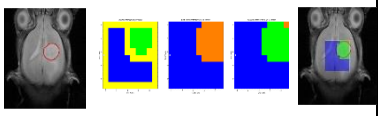
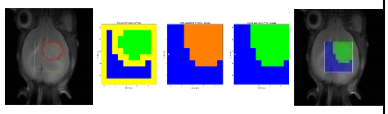
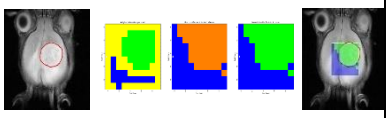
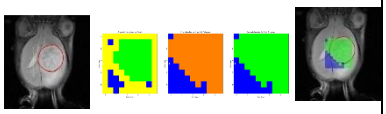
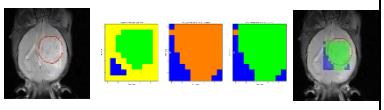
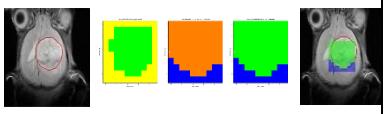
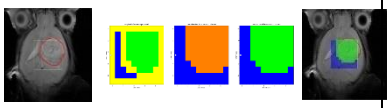
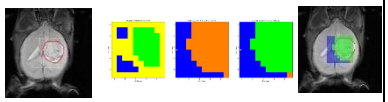
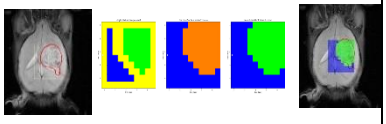
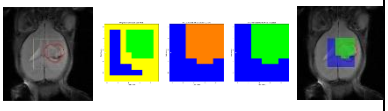
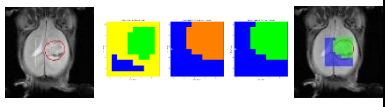
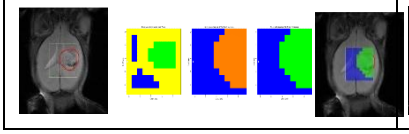
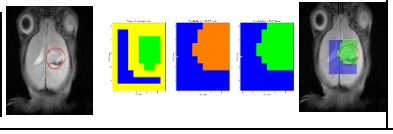
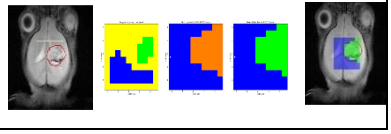
Consecuentemente, aquellos datos que fueron identificados como tumor son evaluados por el segundo modelo de 1D CNN. Aquí se clasifica con el color naranja aquellas zonas que son identificadas como tumor sin tratamiento (T) y con color verde aquellas que presentan tratamiento (R).

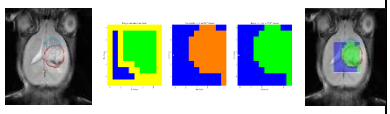
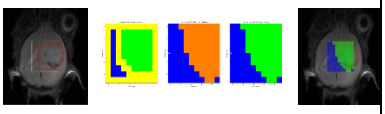
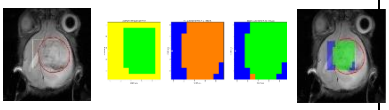
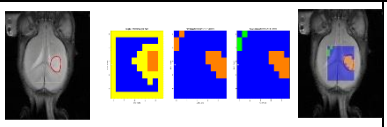
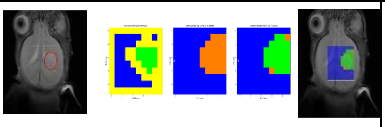
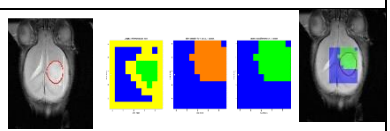
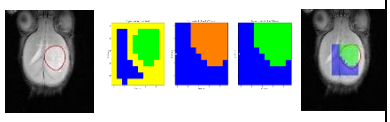
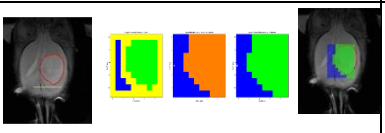
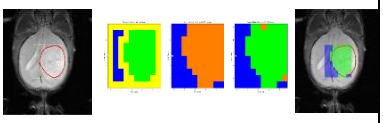
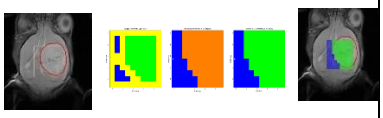
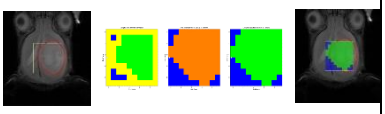
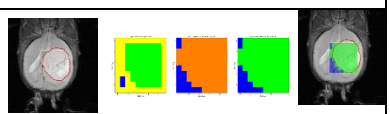
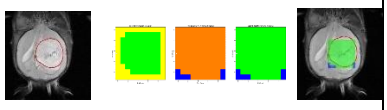
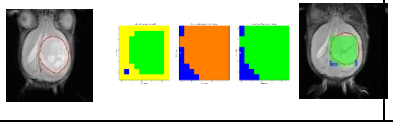
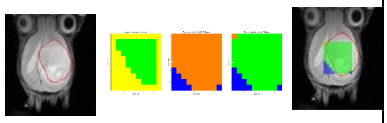
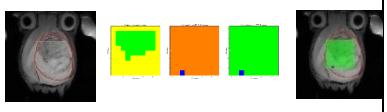
Teniendo en cuenta que cada señal, además de tener una clasificación por un experto, tienen una posición de acuerdo con un plano X y Y que forman cada vóxel y al final la figura 10 x 10 que se muestra a continuación; se asigna una predicción a cada posición del plano y con esta información se realiza el mapa, que, al superponerlo sobre la imagen original que delimita la zona de análisis, permite observar la delimitación de zonas sanas y enfermas por medio de los colores designados. De esta manera, se puede observar el contraste de la predicción y los datos actuales, además de presentar una posible clasificación para aquellas zonas que hasta el momento eran desconocidas (X).

Considerando que el grupo A es de control y los tumores no tienen tratamiento, se esperaría que la clasificación permaneciera en dos clases, N y T. Sin embargo, se puede apreciar que en algunos ratones aparecen vóxeles identificados como tumor en respuesta (R), esto es posible dado que la diferencia entre clases R y T es un problema complicado de ejecutar y en el modelo escogido siguen presentándose (aunque pocos) casos en donde una zona sin tratamiento sea vista como con tratamiento.

A pesar de esta situación, se observa que el modelo realiza un excelente trabajo en contraste con la clasificación actual, respetando aquellas zonas que se encuentran normales y aquellas que están enfermas. La ventaja que ofrece este modelo es permitir la delimitación entre las zonas y poder identificar el punto en el cual empieza el tejido tumoral. Los resultados de este ejercicio se pueden observar con mayor claridad en la figura 14

Figura 14 Mapa de clasificación para aprendizaje semi supervisado

GRUPO A (Todos los sujetos de estudio)		
Ratón C69	Ratón C32	Ratón C71
		
Ratón C179	Ratón C233	Ratón C234
		
Ratón C278		
		
GRUPO C (Sujeto de estudio C819)		
Ratón C819 Día 10	Ratón C819 Día 12	Ratón C819 Día 14
		
Ratón C819 Día 16	Ratón C819 Día 18	Ratón C819 Día 20
		
Ratón C819 Día 22	Ratón C819 Día 24	Ratón C819 Día 26
		
Ratón C819 Día 28	Ratón C819 Día 30	Ratón C819 Día 32
		
Ratón C819 Día 34	Ratón C819 Día 36	Ratón C819 Día 39
		

Ratón C819 Día 41	Ratón C819 Día 43	Ratón C819 Día 45
		
GRUPO C (Sujeto de estudio C821)		
Ratón C821 Día 10	Ratón C821 Día 12	Ratón C821 Día 14
		
Ratón C821 Día 16	Ratón C821 Día 18	Ratón C821 Día 20
		
Ratón C821 Día 22	Ratón C821 Día 24	Ratón C821 Día 26
		
Ratón C821 Día 28	Ratón C821 Día 30	Ratón C821 Día 32
		
Ratón C821 Día 34		
		

3.9 Modelo de máquinas de soporte vectorial para problemas de clasificación de tejidos multiclase – SVM

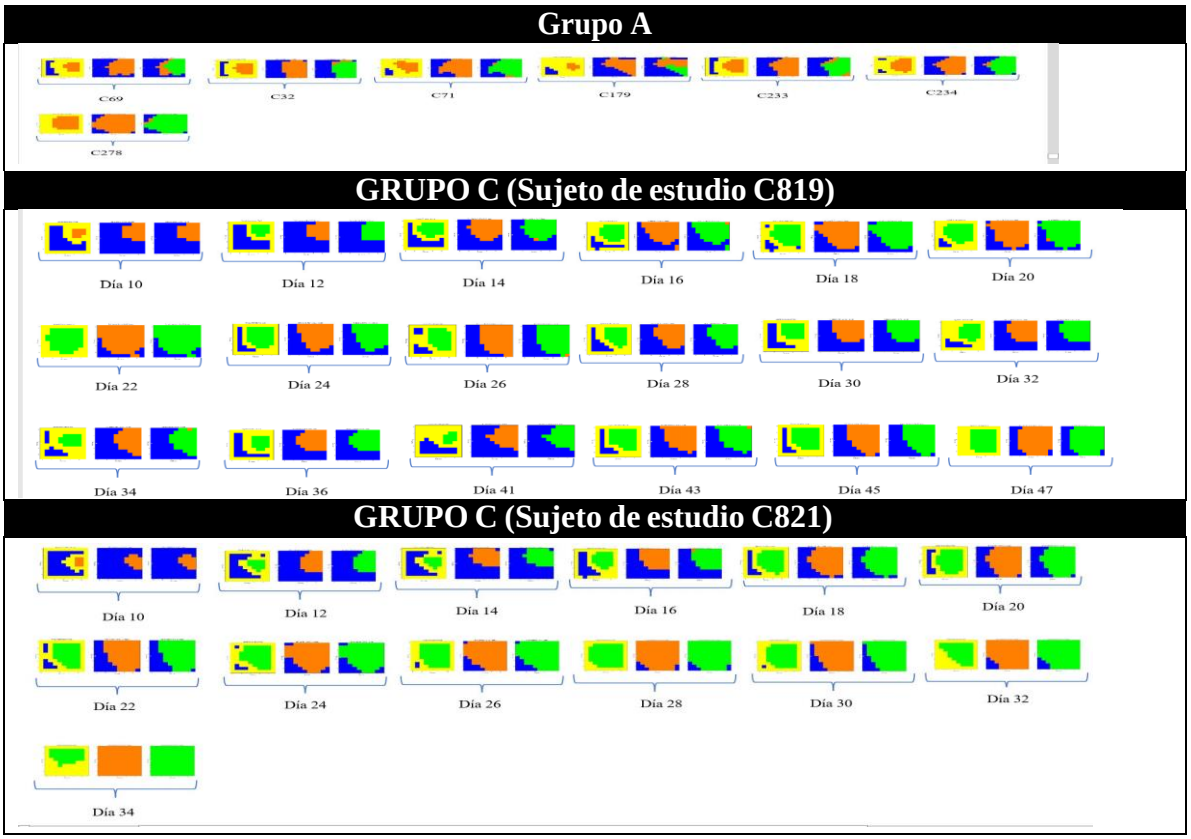
Los modelos expuestos anteriormente fueron el resultado de la prueba de distintos modelos que dado su desempeño fueron desestimados para el objeto demostrativo dentro de este trabajo. Sin embargo, se debe considerar su creación puesto que hicieron parte de todo el desarrollo y permitieron el hallazgo de los resultados finales expuestos.

Con respecto a los modelos tradicionales de machine learning se plantearon como primeros modelos aquellos cuyos parámetros se encuentran dados por el algoritmo de sklearn, esto con

el objetivo de poder evidenciar su desempeño en un primer escenario. Posteriormente, se ejecutaron distintas técnicas de optimización de hipeparámetros que dieron como resultado una variedad de modelos con métricas distintas lo que condujo al establecimiento del mejor modelo por cada algoritmo de aprendizaje.

Referente a los modelos de redes convolucionadas de una dimensión (1D CNN), se destaca el modelo anterior al modelo definitivo de redes cuyas predicciones no resultaron siendo óptimas. Un ejemplo de algunos de los resultados obtenidos se puede visualizar en la figura 15.

Figura 15 Mapa de clasificación para aprendizaje semi supervisado de modelo 1D CNN previo al definitivo



÷

Este modelo demostró resultados ineficientes a pesar de contar con la misma estructura, debido a que el número de épocas de ejecución era menor y no lograba converger hacia el modelo con mejores resultados. Adicionalmente, no contaba con la técnica de *early stopping* que significó la presencia de sobreajuste de los datos de entrenamiento y que la función de pérdida fuese mayor de lo que debería. Al corregir esto, se encontró un modelo que permitió clasificar correctamente las clases.

4. Evaluación y comparación de los modelos establecidos para la clasificación de los tejidos cerebrales comprometidos con GBM

Finalmente se tienen 10 modelos para evaluación, dos por cada algoritmo planteado a lo largo del trabajo de investigación. Los dos modelos desarrollados por cada algoritmo cumplen funciones distintas. El primer modelo deberá detectar y clasificar si existe presencia de tumor sobre la zona a analizar: RT o N. El segundo modelo deberá ser capaz de establecer cuando una zona tumoral ha recibido tratamiento y cuando no: R o T.

4.1 Aplicación de la evaluación

A continuación, se analizarán los resultados del primer modelo propuesto con cada algoritmo de clasificación y se resumirán las métricas de evaluación escogidas para este ejercicio en la tabla 16.

Tabla 16 Métricas de evaluación para los algoritmos de clasificación del modelo 1

Modelo 1 N vs RT					
Models	Accuracy	Recall	Specificity	Precision	F1-Score
Logistic Regresion	0.965	0.970	0.879	0.994	0.977
SVM	0.977	0.977	0.923	0.994	0.985
Random Forest	0.979	0.977	0.923	0.997	0.987
XGBoost	0.969	0.965	0.889	0.995	0.980
1D CNN	0.993	0.993	0.986	0.997	0.995

Como se puede observar en la tabla 16 donde se resumen las métricas del modelo 1, entre los modelos tradicionales de ML, el modelo que presenta una mayor tasa de predicciones correctas es el modelo de Random Forest, seguido por el modelo SVM. En la misma línea, se evidencia que, para el objetivo de clasificar correctamente los casos con tumor, el desempeño de este modelo es el más sobresaliente.

Sin embargo, si contrastamos estos resultados con los del modelo de redes convolucionadas de una dimensión (1D CNN) podemos observar que la exactitud de este modelo es mayor. Igualmente, la capacidad que tiene el modelo para detectar una zona con tumor (RT) y una zona sana (N) mejora bastante, pasando de un índice de especificidad de 0.923 (random forest) a 0.986 (1D CNN) y de un índice de sensibilidad de 0.977 (random forest) a 0.993 (1D CNN).

Esto quiere decir que el último modelo desarrollado es el que menos casos falsos negativos y falsos positivos genera, lo cual es el objetivo principal en este problema. En otras palabras, la estructuración de un modelo que reduzca al máximo la posibilidad de diagnosticar un tejido enfermo como sano o un tejido sano como enfermo.

Con respecto a los modelos de ML tradicionales, se evidencia que el encontrar un punto de corte óptimo benefició específicamente la métrica de especificad, luego se buscaba encontrar el punto en donde la sensibilidad y la especificidad encontrarán un balance, permitiendo efectuar un modelo con mayor robustez.

En cuanto el segundo modelo aplicado para clasificar un tumor con tratamiento (R) de otro sin tratamiento (T), se obtuvieron los siguientes resultados, resumidos en la tabla 17:

Tabla 17 Métricas de evaluación para los algoritmos de clasificación del modelo 2

Modelo 2 R vs T

Models	Accuracy	Recall	Specificity	Precision	F1-Score
Logistic Regresion	0.921	0.933	0.665	0.973	0.953
SVM	0.922	0.929	0.680	0.980	0.953
Random Forest	0.845	0.832	0.484	0.983	0.902
XGBoost	0.886	0.890	0.573	0.975	0.930
1D CNN	0.997	0.998	0.986	0.999	0.999

En este segundo modelo, el esfuerzo es mayor en relación con el primero, debido a la semejanza de las características entre la clase R y T. En general se aprecia un mejor desempeño de los modelos lineales de ML en comparación con los modelos no lineales. Lo anterior se detecta principalmente al observar la métrica de especificidad, la cual en random forest y en XG-Boost es cercana al 0.5, demostrando su incapacidad para reconocer efectivamente las observaciones sin presencia de tumor. La magnitud de casos que son diagnosticados como enfermos cuando su estado real de salud es un tejido sano, es mayor en estos dos modelos en comparación con los otros.

Con respecto a los modelos lineales, se evidencia que a pesar de que las métricas como exactitud y sensibilidad muestran mejores resultados, la especificidad sigue siendo muy baja. En el caso de máquinas de soporte vectorial (SVM). No obstante, en términos de exactitud y precisión este es el modelo con mejores indicadores. Al observar la especificidad, se puede ver que la probabilidad de que el resultado de la predicción sea tumor si la zona tumoral no estuviese tratada, es del 68%.

El balance entre sensibilidad y especificidad nos permiten determinar qué tan confiables podrán ser los resultados del modelo. En este sentido, de igual modo que con el modelo 1, las redes convolucionadas de una dimensión reflejan un excelente desempeño en la clasificación correcta de la presencia de tratamiento o no en una zona del cerebro con un tumor.

El modelo 2 tiene una particularidad y es la diferencia en la cantidad de la muestra que existe en ambas clases. El 88% de la información disponible para construir este modelo (datos de entrenamiento y testeo) pertenece a la clase R, mientras que la clase T posee el 12% de la información. Por esta razón, fiarse solamente de la precisión o de la exactitud podría no ser lo

más correcto en esta situación puesto que se basan principalmente en la clasificación de la clase positiva (R). De esta manera el F1-Score permite capturar los rasgos asociados a ambos tipos de errores al balancear el indicador incluyendo la sensibilidad para su obtención. En este aspecto se observa que los modelos con un mayor F1-Score son los modelos lineales y la red convolucionada de una dimensión (1D CNN).

5 Conclusiones y Recomendaciones

Este trabajo se fundamenta en la detección de la presencia de tumor en una zona específica del cerebro, y adicionalmente busca señalar cuando un tumor se encuentra en tratamiento o no. Para alcanzar este objetivo se planteó la metodología KDD, por considerarla la más pertinente en el estudio de este tipo de investigación.

Dando respuesta al primer objetivo específico, en la selección de datos se trabajó con información que corresponde al conjunto de señales provenientes de las imágenes MRSI de distintos individuos (ratones) que hicieron parte de un estudio del Departamento de Bioquímica y Biología Molecular y la Unidad de Patología Murina y Comparada del Departamento de Medicina y Cirugía Animal, ambos de la Universidad Autónoma de Barcelona en España, sobre el GBM.

Continuando con el segundo objetivo, se construyó el pipeline que unificaría los datos en cuestión con el objetivo de posteriormente realizar su preprocesamiento, transformación, entrenamiento y evaluación de los modelos desarrollados.

Al preprocesar la información no se encontraron datos atípicos y se identificaron las etiquetas N, R, T y X que direccionaron la investigación. Luego, en la etapa de transformación de los datos se pudo consolidar la información, al separar los espectros por ficheros de señales, posiciones y adicional por grupo de estudio de los ratones. Se concluyó que era necesario eliminar la clasificación X, pues al poseer la mayor cantidad de registros desbalanceaba la distribución de los datos y presentaba comportamientos de los metabolitos como aquellos de las otras tres clases, es decir no poseía una etiqueta clara para la aplicación de los modelos.

Posteriormente se procedió al diseño de los modelos de predicción sobre la clasificación de tejidos cerebrales comprometidos con GBM. Se definieron dos enfoques principales para dar solución a esta problemática, en primer lugar, hacer uso de modelos de ML como regresión logística, máquinas de soporte vectorial, random forest y XG-Boost; en segundo lugar, utilizar DL con redes convolucionadas de una dimensión, aprovechando las bondades de este tipo de algoritmos sobre los espectros derivados de MRSI. Se empleó como herramienta principal para los modelos de ML la librería de Sklearn, y para DL, tensorflow.

De esta manera, la contribución principal de este trabajo es la evaluación de una nueva metodología que permita la clasificación de tejidos normales (N) y con tumor (RT), a través de métodos de ML y DL, siendo las redes convolucionadas de una dimensión (1D CNN) las que logran mejores resultados. Esto permite detectar oportunamente el crecimiento de un tejido tumoral. Aprovechando la información adicional que generan los espectros al captar el comportamiento de los metabolitos, se puede solventar la falta de precisión que pudiesen tener otros métodos no invasivos para visualizar el estado del tejido cerebral, como MRSB, trabajados en las investigaciones anteriores.

Para este objetivo, se concluyó que era necesario efectuar dos modelos de clasificación luego de realizar un ejercicio de visualización usando las técnicas PCA y TSNE, encontrando que la clase N se distinguía bastante bien de las clases R y T. Sin embargo, entre estas dos últimas resultó más compleja su diferenciación y se evidenció que estas se componían de vectores similares. De este modo, se propuso un primer modelo para identificar la clase N de las clases RT, y un segundo modelo para reconocer entre R y T. Se presentaron dos grandes desafíos para este último propósito: (i) el hecho de que la muestra se encontrará desbalanceada implicando un sobreajuste sobre la clase mayoritaria (R) y, (ii) el hecho de que ambas clases involucran tumor, dificultando la segmentación de características.

Con el objetivo de corregir los problemas intrínsecos del desbalanceo de la data se concluyó que era necesario utilizar técnicas de sobre muestreo (oversampling) con la técnica estadística SMOTE. Adicionalmente, para los modelos de ML se calculó un punto de corte óptimo que definiera un umbral de clasificación, pues de esto dependerá el desempeño del modelo medido en precisión, sensibilidad y especificidad. Ahora bien, para los modelos de DL se utilizaron técnicas como el uso de un optimizador Adam pretendiendo minimizar la función de pérdida y determinando el desempeño del modelo. Además, se empleó una técnica de regularización (early stopping) para asegurar que el modelo que se escogiera fuera el mejor posible.

En el desarrollo de los modelos se descartó el de redes neuronales para problemas de clasificación de tejidos multiclase (2D CNN), pues se encontró que el desbalance de las clases (N, R y T). al asemejar la data de las señales a imágenes 2D no permitía que el modelo convergiera a una solución del problema planteado en la investigación, a pesar de aplicar técnicas de sobremuestreo buscando balancear las categorías de clasificación.

Finalmente, el último objetivo se enfoca en la comparación de estos modelos y la evaluación realizada, que tuvieron presentes métricas relacionadas con estudios médicos, analizando el desempeño de los modelos por medio de la especificidad de los resultados y así poder evaluar la capacidad de discriminar los falsos negativos que se traduce en la falta de la detección temprana de la enfermedad (Komori, 2022).

A partir de la evaluación de los resultados de cada modelo desarrollado y expuesto en la investigación, se concluyó que, para el primer modelo, desarrollado con el fin de discriminar entre la clase normal (N) y las clases con tumor (RT), las redes convolucionadas de una dimensión (1D CNN) superan otros clasificadores. Se demostró que la técnica de DL aprovecha las características de los espectros para poder identificar de la mejor forma posible las diferencias entre las características propias de cada clasificación. No obstante, también se concluye que, en términos generales, los modelos desarrollados de ML demuestran un buen desempeño.

Cuando se revisa el segundo modelo que tiene como objetivo discernir entre clase de tumor con respuesta (R) y sin respuesta (T), el escenario es diferente. En tales circunstancias se evidencian las bondades de las redes convolucionadas de una dimensión (1D CNN) para reconocer la evolución del espectro como señal y no como variables individuales. En otras

palabras, DL resulta más útil y aporta un mejor resultado que los modelos lineales y no lineales de ML. Esto es relevante puesto que, la clasificación actual generada por los expertos no permite identificar tan fácilmente las peculiaridades que disocian ambas clases.

Así las cosas, esta investigación presentó la comparación entre varios modelos de aprendizaje automático con el fin de establecer el mejor modelo para predecir si hay respuesta a la terapia en pacientes comprometidos con GBM, utilizando señales derivadas de MRSI. Se concluye que la aplicación de modelos de clasificación lineales, no lineales y redes neuronales convolucionadas de una dimensión permiten determinar la temprana identificación del nivel de respuesta a la terapia, con el objetivo de personalizar los tratamientos para el GBM y así mejorar su eficacia, pues podría incrementar la tasa de supervivencia de los pacientes.

A partir del ejercicio realizado en la investigación, se pudo mejorar la interpretación de los datos obtenidos en la fase de tratamiento, para ayudar a entender, de una manera no invasiva, si los tumores están en respuesta a la terapia, a partir de la composición química de las muestras revelando la información metabólica (biomarcadores) (Horská & Barker, 2010).

Se concluye que la respuesta al problema de investigación es positiva, toda vez que ***si es posible establecer un modelo de alto desempeño para predecir a qué tipo de clase pertenecen los tejidos del cerebro comprometidos con GBM, analizando las señales derivadas de imágenes espectroscópicas de resonancia magnética (MRSI), con la aplicación de modelos de clasificación lineales, no lineales y redes neuronales convolucionadas de una dimensión, con el fin de determinar si existe respuesta al tratamiento suministrado.***

Concretamente, el uso de 1D CNN demostró un excelente desempeño para predecir a qué tipo de clase pertenecen los tejidos del cerebro comprometidos con GBM. Es decir que, con esta herramienta fue posible evaluar el comportamiento del tumor al delimitar la zona en donde se encuentra y ver su evolución a través del tiempo, permitiendo evaluar oportunamente su condición y que sea posible tomar decisiones para el tratamiento más efectivo de la enfermedad.

En el futuro se recomienda realizar distintas combinaciones de separación de muestra entre entrenamiento y testeo con técnicas de validación cruzada con el objetivo de descartar problemas de sobreajuste o subajuste en los modelos lineales y no lineales y así poder generar predicciones más certeras. Igualmente se podrían considerar distintas técnicas de balanceo de data y debería contemplarse la posibilidad de combinar distintos algoritmos de aprendizaje para el modelo 1 y el modelo 2.

Bibliografía

Association of Swedish Automobile Manufacturers and Wholesalers. (1997). *Producentansvar för uttjänta bilar: Rapport till Naturvårdsverket om hur bilindustrin avser att hantera producentansvaret för uttjänta bilar. [Producer responsibility for used cars: Report to the Environmental Protection Agency on how the car industry intends to handle the producer responsibility for used cars]*. 1 October 1997. Stockholm: Bilindustriföreningen.

Backman, Mikael, Huisingh, Donald, Lidgren, Karl, & Lindhqvist, Thomas. (1988). *Om en avfallsstyrd produktutveckling [About a Waste Conscious Product Development]*. Report 3488. Solna: Swedish Environmental Protection Agency.

Brinkmann, Walter, & Fonteyne, Jacques. (1999). Extended Producer Responsibility. Monitoring Performance. In *OECD Workshop on Extended Producer Responsibility and Waste Minimization Policy in Support of Environmental Sustainability*, 4-7 May 1999, Paris.

Balamurugan, T., & Gnanamanoharan, E. (2022). *Brain Tumor Segmentation and Classification using hybrid Deep CNN with LuNet Classifier* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-1599383/v1

Bantis, L. E., Nakas, C. T., & Reiser, B. (2014). Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point: Construction of Confidence Regions in the ROC Space after the Estimation of the Cut-Off Point. *Biometrics*, *70*(1), 212–223. https://doi.org/10.1111/biom.12107

Berrar, D. (2019). Performance Measures for Binary Classification. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 546–560). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20351-8

Boytsov, A., Fouquet, F., Hartmann, T., & LeTraon, Y. (2017). *Visualizing and Exploring Dynamic High-Dimensional Datasets with LION-tSNE* (arXiv:1708.04983). arXiv. http://arxiv.org/abs/1708.04983

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Delgado-Goñi, T., Ortega-Martorell, S., Ciezka, M., Olier, I., Candiota, A. P., Julià-Sapé, M., Fernández, F., Pumarola, M., Lisboa, P. J., & Arús, C. (2016). MRSI-based molecular imaging of therapy response to temozolomide in preclinical glioblastoma using source analysis: Glioblastoma Therapy Response Detection by MRSI and Source Analysis. *NMR in Biomedicine*, *29*(6), 732–743. https://doi.org/10.1002/nbm.3521

Dolecek, T. A., Propp, J. M., Stroup, N. E., & Kruchko, C. (2012). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2005-2009. *Neuro-Oncology*, *14*(suppl 5), v1–v49. https://doi.org/10.1093/neuonc/nos218

Gopika, N., & kowshalaya M.E., A. M. (2018). Correlation Based Feature Selection Algorithm for Machine Learning. *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, 692–695. https://doi.org/10.1109/CESYS.2018.8723980

Govindaraju, V., Young, K., & Maudsley, A. A. (2000). Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR in Biomedicine*, *13*(3), 129–153. https://doi.org/10.1002/1099-1492(200005)13:3<129::aid-nbm619>3.0.co;2-v

Gupta, L. (2020, November). *Comparison of Hyperparameter Tuning algorithms: Grid search, Random search, Bayesian optimization*. https://medium.com/analytics-vidhya/comparison-of-hyperparameter-tuning-algorithms-grid-search-random-search-bayesian-optimization-5326aaef1bd1

Horská, A., & Barker, P. B. (2010). Imaging of Brain Tumors: MR Spectroscopy and Metabolic Imaging. *Neuroimaging Clinics of North America*, *20*(3), 293–310. https://doi.org/10.1016/j.nic.2010.04.003

Hortúa, H. J. (2022, Primer semestre). *¨Hello world con Tensorflow 2. Una introducción a Deep Learning¨. Actividad #7.*

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398. https://doi.org/10.1016/j.ymssp.2020.107398

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, *1*(3).

Koehrsen, W. (2018, January). *Hyperparameter Tuning the Random Forest in Python.* https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Komori, T. (2022). [The 2021 WHO Classification of Tumors, 5th edition, Central Nervous System Tumors: A Short Review]. *Brain and Nerve = Shinkei Kenkyu No Shinpo*, *74*(6), 803–809. https://doi.org/10.11477/mf.1416202124

KOPOT, A. (n.d.). NMR SPECTROSCOPY PLAYLIST: Parts Per Million in NMR Spectroscopy. *Medical and Science*. https://aklectures.com/lecture/parts-per-million-in-nmr-spectroscopy

Kouli, O., Hassane, A., Badran, D., Kouli, T., Hossain-Ibrahim, K., & Steele, J. D. (2022). Automated brain tumor identification using magnetic resonance imaging: A systematic review and meta-analysis. *Neuro-Oncology Advances*, *4*(1), vdac081. https://doi.org/10.1093/noajnl/vdac081

Luts, J., Heerschap, A., Suykens, J. A. K., & Van Huffel, S. (2007). A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection. *Artificial Intelligence in Medicine*, *40*(2), 87–102. https://doi.org/10.1016/j.artmed.2007.02.002

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and others. (2016).OSDI , 16, 265-283. tensorflow.org.

Mbaabu, O. (2022, December). *Introduction to Random Forest in Machine Learning*. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Moreno, R. A., & Holodny, A. I. (2021). Functional Brain Anatomy. *Neuroimaging Clinics of North America*, *31*(1), 33–51. https://doi.org/10.1016/j.nic.2020.09.008

Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

Ortega-Martorell, S., Candiota, A. P., Thomson, R., Riley, P., Julia-Sape, M., & Olier, I. (2019). Embedding MRI information into MRSI data source extraction improves brain tumour delineation in animal models. *PLOS ONE*, *14*(8), e0220809. https://doi.org/10.1371/journal.pone.0220809

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Parker, N. R., Khong, P., Parkinson, J. F., Howell, V. M., & Wheeler, H. R. (2015). Molecular Heterogeneity in Glioblastoma: Potential Clinical Implications. *Frontiers in Oncology*, *5*. https://doi.org/10.3389/fonc.2015.00055

Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data* (First edition). O'Reilly Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Prechelt, L. (2012). Early Stopping—But When? In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (Vol. 7700, pp. 53–67). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_5

Premanand, S. (2021, July 27). Convolution Neural Network – CNN Illustrated With 1-D ECG signal. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/07/convolution-neural-network-the-base-for-many-deep-learning-algorithms-cnn-illustrated-by-1-d-ecg-signal-physionet/

Shultz, T. R., Fahlman, S. E., Craw, S., Andritsos, P., Tsaparas, P., Silva, R., Drummond, C., Ling, C. X., Sheng, V. S., Drummond, C., Lanzi, P. L., Gama, J., Wiegand, R. P., Sen, P., Namata, G., Bilgic, M., Getoor, L., He, J., Jain, S., … Mueen, A. (2011). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 209–209). Springer US. https://doi.org/10.1007/978-0-387-30164-8_157

Support Vector Machine (SVM) Hyperparameter Tuning In Python. (2022, May). https://medium.com/grabngoinfo/support-vector-machine-svm-hyperparameter-tuning-in-python-a65586289bcb

Tan, A. C., Ashley, D. M., López, G. Y., Malinzak, M., Friedman, H. S., & Khasraw, M. (2020). Management of glioblastoma: State of the art and future directions. *CA: A Cancer Journal for Clinicians*, *70*(4), 299–312. https://doi.org/10.3322/caac.21613

Tuhin, Md. A. H., Pramanick, T., Emon, H. K., Rahman, W., Rahi, Md. M. I., & Alam, Md. A. (2020). Detection and 3D Visualization of Brain Tumor using Deep Learning and Polynomial Interpolation. *2020 IEEE Asia-Pacific*

Conference on Computer Science and Data Engineering (CSDE), 1–6. https://doi.org/10.1109/CSDE50874.2020.9411595

Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/cait-2019-0001

Verma, S. (2019, September). Understanding 1D and 3D Convolution Neural Network | Keras. *Towards Data Science*. https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

Wolburg, H., Noell, S., Fallier-Becker, P., Mack, A. F., & Wolburg-Buchholz, K. (2012). The disturbed blood–brain barrier in human glioblastoma. *Molecular Aspects of Medicine*, *33*(5–6), 579–589. https://doi.org/10.1016/j.mam.2012.02.003

Yin, J., & Tian, L. (2014). Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Computational Statistics & Data Analysis*, *77*, 1–13. https://doi.org/10.1016/j.csda.2014.01.021

Zhu, T., Lin, Y., & Liu, Y. (2017). Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognition*, *72*, 327–340. https://doi.org/10.1016/j.patcog.2017.07.024

Zhu, X. P., Young, K., Ebel, A., Soher, B. J., Kaiser, L., Matson, G., Weiner, W. M., & Schuff, N. (2006). Robust analysis of short echo time1H MRSI of human brain. *Magnetic Resonance in Medicine*, *55*(3), 706–711. https://doi.org/10.1002/mrm.20805

Zhang, Haotian, Lin Zhang, y Yuan Jiang. «Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems». En *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 1-6. Xi'an, China: IEEE, 2019. https://doi.org/10.1109/WCSP.2019.8927876.

Abreviaciones

1D-CNN	Redes convolucionadas de una dimensión
2D CNN	Redes convolucionadas de dos dimensiones
DL	Aprendizaje profundo (Deep Learning)
EPA	Agencia para la protección Ambiental (Environmental Protection Agency)
GBM	Glioblastoma
ML	Aprendizaje automático (Machine learning)
MRI	Imágenes por resonancia magnética (Magnetic Resonance Imaging)
MRSI	Imágenes espectroscópicas por resonancia mangnética (Magnetic Resonance Spectroscopic Imaging)
PCA	Técnica matemática de análisis de componentes principales
PET	Tomografía de emisión de positrones
ROC AUC	Área bajo la curva ROC (Receiver Operating Characteristic of the Area Under the Curve)
SMOTE	Xxx (Synthetic Minority Oversampling Technique))
STE	Tiempo de eco corto (Short Time of Eco)
SVM	Modelo de máquinas de soporte vectorial (Support-vector machines)
TE	Tiempo de eco (Time Eco)
t-SNE	Técnica probabilística de implante de vecinos estocásticos t-Diseminados (T-distributed Stochastic Neighbor Embedding)
XGBOOST	Refuerzo de gradientes extremo (Extreme Gradient Boosting)