Alba Giró Subirats

Radioteràpia ablativa amb intenció curativa pel tractament de pacients amb càncer de pulmó: incorporació de la metabolòmica per assistir en les decisions clíniques amb models d'intel·ligència artificial

TREBALL DE FI DE GRAU

Dirigit per Dra.D Meritxell Arenas Prat i Dr. Victor Hernández Masgrau

Grau d'Enginyeria biomèdica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2024

Als meus pares, per estar sempre al meu costat en aquesta etapa i en totes les que han passat i les que vindran.

Al Víctor i la Meritxell, per oferir-me l'oportunitat de treballar en un projecte tan apassionant i enriquidor, que m'ha inspirat a endinsar-me en el món de les dades clíniques.

Resum

Aquest treball explora l'ús de la radioteràpia ablativa amb intenció curativa per al tractament del càncer de pulmó, incorporant tècniques de metabolòmica per enriquir la presa de decisions clíniques mitjançant models d'intel·ligència artificial. S'examina la relació entre dades clíniques, radiòmiques i metabolòmiques per identificar les variables més significatives per a l'entrenament dels models, millorant així la seva eficiència. A partir d'aquesta selecció de dades, s'implementen dos arquitectures d'intel·ligència artificial per predir els resultats del tractament. Els resultats subratllen el potencial d'integrar múltiples fonts de dades per millorar les decisions clíniques, oferint una nova perspectiva per a futures recerques i aportant informació prospectiva per a la presa de decisions clíniques. Aquest enfocament promet avenços significatius en la precisió dels diagnòstics i en la personalització dels tractaments, marcant un progrés cap a la medicina de precisió i personalitzada en l'oncologia pulmonar.

Abstract

This study investigates the application of ablative radiotherapy with curative intent in lung cancer treatment, using metabolomics techniques to enhance clinical decision-making through artificial intelligence models. It analyzes the relationship between clinical, radiomic and metabolomics data to use the most significant in training models, thereby improving their efficiency. From this selection, two artificial intelligence models are implemented to predict treatment outcomes. The results highlight the potential of combining multiple data sources to enhance clinical decisions, offering a new perspective for future studies and providing prospective information for clinical decision-making. This approach promises significant improvements in diagnostic accuracy and treatment personalization, marking a step toward precision and personalized medicine in pulmonary oncology.

Índex

Índex de figures

1. Introducció

1.1 Contextualització del càncer de pulmó.

Es defineix el càncer com el resultat d'aquell procés que provoca un creixement i divisió incontrolat de les cèl·lules. Les cèl·lules que posseeixen aquestes característiques s'anomenen cèl·lules cancerígenes i poden formar tumors, els quals poden provocar un mal funcionament de l'òrgan en el que es troben i conseqüentment un mal funcionament del cos[1].

Segons la base de dades *Global Cancer Observatory* creada per la *International Agency of Research on Cancer* (IARC), l'any 2022 el 18,7% de les defuncions relacionades amb el càncer van ser causades pel càncer de pulmó.

Si diferenciem per sexe, el càncer de pulmó és el més diagnosticat i el més letal per el que fa als homes. Mentre que per el que fa a les dones és el tercer[2].

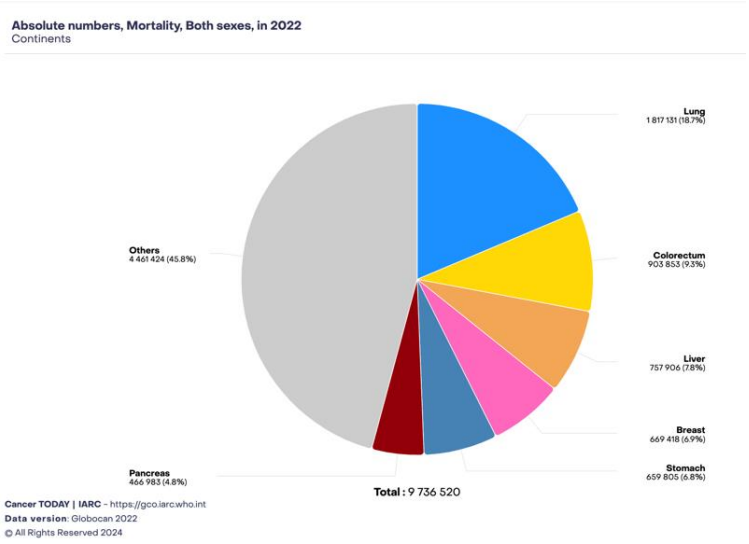


Figura 1. Gràfic circular que representa el percentatge de mortalitat, a nivell mundial, de diferents càncers.

Com s'ha demostrat, el càncer de pulmó és un dels més letals a nivell mundial, sent la taxa de supervivència després dels 5 anys d'un 15% aproximadament[3].

Pel que fa al diagnòstic de pacients amb sospita de càncer de pulmó, s'inclou un estudi tissular a partir d'una mostra histològica. Aquest dona informació sobre l'estadi l'estat metastàtic a part d'una avaluació funcional del pacient. Tot i que l'estudi metastàtic es complementa també amb dades clíniques, tomografies computeritzades i tomografia per emissió de positrons, entre d'altres.[3]

1.2 Tipus de càncer de pulmó, histologia i estadis

Els càncers de pulmó es classifiquen en dos grups diferents carcinoma de cèl·lules petites (SCLC) o carcinoma de cèl·lules no petites (NSCLC). Aquesta classificació s'utilitza per a prendre referents al diagnòstic i també per determinar els pronòstic de la malaltia[3].

1.2.1 Carcinoma de cèl·lules petites

Per al que fa a la seva histologia, són carcinomes microcítics formats per cèl·lules petites. Aquestes són de forma ovalada i amb citoplasma petit. En les imatges radiològiques es presenten com fils en les zones dels bronquis.

Figura 2. Imatge radiològica de pacient amb SCLC

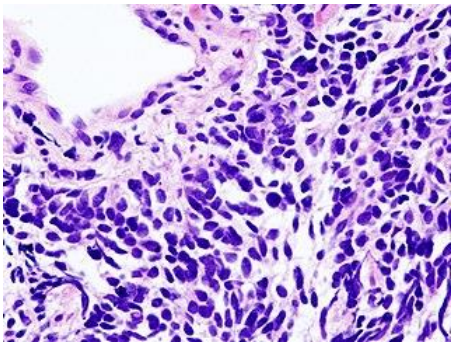


Figura 3. Imatge histològica de pacient amb SCLC

El SCLC representa el 15% dels carcinomes broncogènics. Es caracteritza per tenir una tendència metastàtica precoç, de manera que la cirurgia té un paper limitat com a teràpia primària. En el cas d'aquests carcinomes, s'opta principalment per quimioteràpia estàndard en combinació amb radioteràpia toràcica. Les respostes a el tractament solen ser altes, tot i que la supervivència a 5 anys és del 15% al 25% per als pacients amb SCLC en estadi limitat i menor de l' 1% per als pacients amb malaltia en estadi extens[4].

1.2.2 Carcinoma de cèl·lules no petites

Representa el 85% de tots els casos de càncer de pulmó. A més, del 85% al 90% d'aquests està causat per fumar. Es caracteritza per tenir un creixement més lent i per tant, té menys probabilitat de metàstasi precoç. Aquest es pot classificar en tres tipus diferents segons la seva histologia i zona d'aparició.

- Adenocarcinoma. És el més comú, s'origina en les cèl·lules que produeixen mucus situades a la ínia exterior dels pulmons. Aquest és més probable que aparegui en persones que han sigut o son fumadores.
- Carcinoma epidermoide. Es solen iniciar en les cèl·lules esquamoses, que són cèl·lules planes que recobreixen l'interior dels conductes aeris en els pulmons. També esta fortament relacionat amb pacients fumadors.
- Carcinoma de grans cèl·lules. Aquest pot aparèixer en qualsevol part del pulmó i tendeix a créixer ràpidament [5].

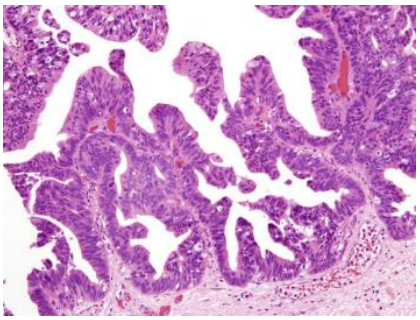


Figura 4. Imatge histològica de pacient amb Adenocarcinoma

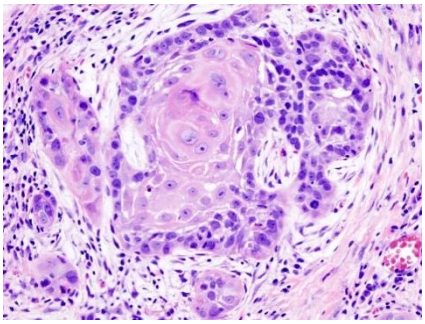


Figura 5. Imatge histològica de pacient amb Carcinoma epidermoide

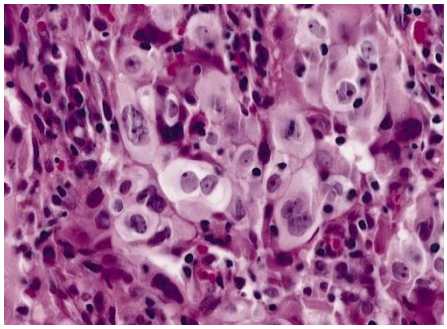


Figura 6. Imatge histològica de pacient amb Carcinoma de cèl·lules grans

1.2.3 Radioterapia estereotàtica corporal (SBRT)

La teràpia de radiació corporal estereotàctica (SBRT) ha sorgit com un estàndard d'atenció a pacients mèdicament inoperables amb càncer de pulmó de cèl·lules no petites en etapes inicials de la malaltia i en situacions oligometastàsiques, on el càncer ha començat a disseminar-se però està en un número petit de llocs [6][7].

Aquesta tècnica es basa en tres principis fonamentals. El primer és la localització estereotàctica precisa, és a adir, utilitzar referències internes o externes per a localitzar de manera precisa. El segon és la imatge diària, que permet localitzar el tumor cada dia, a més de visualitzar els òrgans crítics per assegurar que el tractament s'apliqui de manera correcta i segura. Finalment, el tractament fraccionat, la SBRT s'administra en entre 1 i 5 sessions o fraccions[7].

Tot i que en diferents investigacions parla del risc de toxicitat i efectes secundaris dels tractaments SBRT [1]. Com en el cas dels càncers centrals, com el pulmonar, on presenta més riscs degut a que estan a prop de la tràquea. Tot i així, es demostra també una supervivència general més alta en comparació a altres tècniques de tractament, sobretot en adults grans [6] [7].

1.3 Metabolòmica

1.3.1 Introducció a la metabolòmica

La metabolòmica és una disciplina centrada en l'anàlisi exhaustiu de metabòlits que es troben en les cèl·lules, fluids, teixits i organismes. Aquests metabòlits composen el metaboloma, que és el reflex fisiològic d'un organisme en un instant donat, aportant informació sobre la salut d'aquest identificant canvis en alteracions en rutes metabòliques específiques. D'aquesta manera, la metabolòmic és útil per predir malalties o indicar la resposta a un tractament en concret[8].

Per a poder identificar aquests metabòlits, s'utilitzen plataformes analítiques avançades i tècniques estadístiques de dades amb l'objectiu d'analitzar perfils metabòlics de manera qualitativa i quantitativa. Aquests estudis es poden dividir en estudis dirigits o no dirigits, depenent de si l'estudi es centra en metabòlits específics o en la comparació de la major quantitat d'aquests[8].

1.3.2 Metabolòmica en el càncer de pulmó.

La metabolòmica en el càncer de pulmó es centra en estudiar les alteracions metabòliques en la progressió d'aquesta malaltia. Específicament en el càncer de pulmó NSCLC, s'han identificat

més de cinquanta gens mutats que contribueixen en el creixement i supervivència del tumor[9]. Algunes d'aquestes mutacions es troben en gens com:

- **EGFR** - Receptor del Factor de Creixement Epidèrmic
- **ALK** - Quinasa del Limfoma Anaplàstic
- **ROS1** - Quinasa de la tirosina del proto-oncogen c-ros-1
- **ERBB2** (HER2/neu) - Receptor 2 del Factor de Creixement Epidèrmic Humà
- **MET** - Proto-oncogen c-Met, receptor de la quinasa de la tirosina
- **MAP2K1** - Quinasa 1 de la proteïna activada per mitògens
- **BRAF** - Proto-oncogen B-Raf, quinasa de la serina/treonina
- **KRAS** - Oncogen homòleg del virus del sarcoma de Kirsten en rates
- **NTRK1/2/3** - Receptors de la quinasa de la tirosina de la neurotrofina 1, 2 i 3
- **RET** - Proto-oncogen del receptor de la quinasa de la tirosina del factor de creixement de les cèl·lules de la glàndula tiroidal.

1.4 Radiòmica

1.4.1 Introducció a la Radiòmica

La radiòmica és una ciència que transforma el diagnòstic per la imatge tradicional mitjançant l'ús d'algoritmes automatitzats per estudiar característiques de les imatges mèdiques que són imperceptibles a l'ull humà. Aquesta disciplina vol associar les característiques radiòmiques a estats fisiològics concrets, amb l'objectiu final d'afinar el diagnòstic i el tractament de les malalties, en particular dins l'àmbit de la medicina personalitzada de precisió. La radiòmica mostra un potencial considerable per millorar l'abordatge terapèutic, la investigació clínica i la presa de decisions clíniques basades en un enfoc més individualitzat i precís[10].

En oncologia, la radiòmica pot millorar la precisió en la detecció de tàctiques terapèutiques adequades per als tumors, pot preveure recidives o la aparició de metàstasi, i pot oferir detalls sobre la heterogeneïtat de les lesions que poden passar desapercebudes en una interpretació radiològica estàndard[10].

1.4.2 Radiòmica en el càncer de pulmó.

La radiòmica fa ús de la intel·ligència artificial per transformar imatges mèdiques en dades analitzables, extraient i correlacionant característiques quantitatives de la imatge amb resultats del pacient i el fenotip d'un tumor. La radiòmica s'ha investigat àmpliament en oncologia pulmonar.

El procés d'un estudi radiòmic inclou diversos passos com la recopilació d'imatges mèdiques, la segmentació de l'àrea d'interès, l'extracció de trets radiòmics, la selecció de trets per correlacionar amb resultats d'interès, la construcció de la signatura radiòmica i l'avaluació del rendiment del model. Algunes de les característiques radiòmiques més importants en el càncer de pulmó són[11]:

- **Textura:** Aquestes característiques descriuen la variabilitat, la irregularitat i la complexitat de la intensitat dels píxels dins d'una regió d'interès (ROI), i poden ser indicatives de la heterogeneïtat del tumor.
- **Forma:** Les característiques de la forma del tumor, com la compacitat, la irregularitat de la superfície o l'elongació, poden ajudar a diferenciar entre els tipus de tumors i a correlacionar-los amb els pronòstics.
- **Intensitat:** Mesures com la mitjana, la mediana, o la desviació estàndard de la intensitat de la senyal dins del ROI poden ser útils per a detectar canvis en la densitat dels teixits que poden estar relacionats amb el tipus de teixit tumoral.

- **Histograma:** L'anàlisi del histograma d'intensitat pot revelar la presència de zones de necrosi o hipòxia dins del tumor, les quals poden estar relacionades amb l'agressivitat del tumor i la resposta al tractament.
- **Estadístiques d'ordre superior:** Incloent la correlació, la uniformitat, i l'entropia, que són derivades dels mapes de textura i proporcionen informació sobre la distribució dels valors d'intensitat a l'interior del tumor.
- **Canvis dinàmics:** En l'anàlisi longitudinal, els canvis en les característiques radiòmiques al llarg del temps, com els que es poden calcular amb la δ-radiòmica, poden ser indicatius de la resposta del tumor al tractament o de la seva progressió.
- **Peritumoral:** Les característiques del teixit que envolta el tumor, com ara la presència d'inflamació o canvis en la textura, poden ser indicatius de la invasió tumoral i de la probabilitat de metàstasi.

1.5 Models predictius en decisions clíniques.

La investigació científica en medicina sempre ha buscat establir casualitats per generar intervencions eficaces en contra de les malalties. En els últims anys l'estadística ha jugat un paper molt important en establir relacions entre variables a nivell mèdic.

Gràcies a l'acumulació de grans volums de dades en els registres clínics electrònics i l'increment en el poder computacional, les tècniques d'aprenentatge automàtic s'estan convertint en eines clau en la creació d'anàlisis predictius i reconèixer patrons que no havien sigut identificats prèviament [11]. A més, la capacitat d'aquests models d'interpretar dades no estructurades aporta una gran flexibilitat i dinamisme en la presa de decisions mèdiques.

Per la implementació efectiva de la intel·ligència artificial en salut s'han de considerar aspectes molt importants com la protecció de dades, privacitat i els marcs de normatives nacionals i regionals. Els anàlisis predictius són un dels usos més freqüents dins del "*Machine Learning"*, actualment existeixen molts tipus diferents de models predictius, com xarxes neuronals, màquines de vectors de suport (SVM), regressions logístiques, algoritmes de agrupament i *clustering*, entre d'altres. En aquest projecte específicament, s'han decidit utilitzar dos models amb complexitats diferents. El primer és un model de boscos aleatoris o *Random Forest*, com a model de menor complexitat, i el segon és una xarxa neuronal convolucional o CNN, com a model de major complexitat.

1.5.1 Boscos Aleatoris.

Els models de Random Forest, o Boscos Aleatoris, són una eina d'aprenentatge automàtic molt usats en el món biomèdic per la seva capacitat d'elaborar prediccions precises a partir de grans conjunts de dades complexos. Aquesta tècnica, basada en la creació de múltiples arbres de decisió que treballen conjuntament, destaca per la seva eficiència computacional, la capacitat per manejar dades de gran dimensió i per oferir informació sobre la importància de les variables en la predicció. Per aquests motius es presenta com una opció robusta i relativament accessible per a investigadors i professionals, especialment en camps interdisciplinaris, on simplifica el desenvolupament de models predictius complexos sense requerir un coneixement profund d'estadística, facilitant així l'avanç en la recerca i el desenvolupament de noves solucions biomèdiques[12].

Aquests models operen mitjançant la construcció de múltiples arbres de decisió durant la seva fase d'entrenament. Cada arbre s'entrena amb un subconjunt aleatori del conjunt de dades original, seleccionat amb reposició, en un procés conegut com a mostreig bootstrap. Per realitzar una predicció, Random Forest agrega les decisions de tots els arbres individuals. En tasques de classificació, això es fa per majoria de vots, és a dir, la classe més votada pels arbres és la predicció

del model. En tasques de regressió, es calcula la mitjana de les prediccions de tots els arbres. Aquest enfocament d'agregació ajuda a mitigar el sobreajustament[1] i augmentar la precisió del model[12].

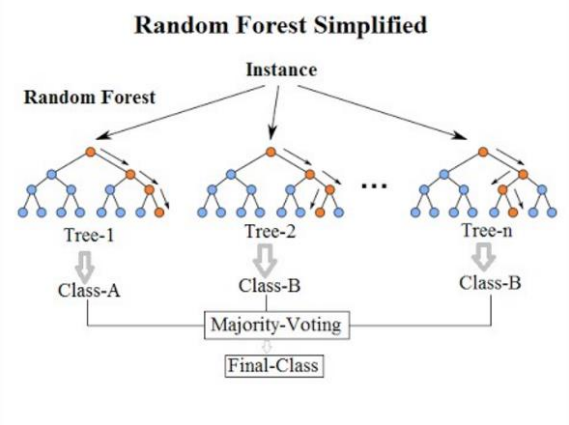


Figura 7. Esquema de funcionament d'un model Random Forest.

1.5.2 Xarxes neuronals

Les xarxes neuronals artificials s'inspiren en les xarxes neuronals humanes. En aquestes, les neurones transmeten senyals elèctrics a altres neurones a través de les dendrites i l'axó. Aquesta interacció entre neurones es produeix a través de la sinapsi. Anàlogament, una neurona artificial consisteix en:

- Sinapsis amb pes i longitud específics.
- Funció de xarxa per sumar senyals d'entrada, generalment amb un combinador lineal.
- Funció d'activació per limitar l'amplitud de la sortida.
- Sortida, que depèn de l'activitat de la neurona.

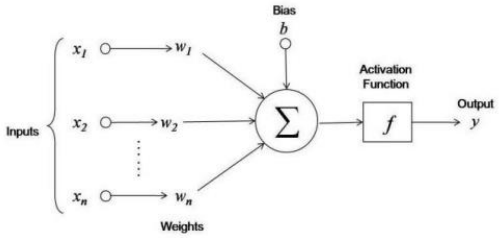


Figura 8. Mode genèric de neurona artificial.

L'arquitectura de les xarxes neuronals pot ser monocapa, on els nodes estan en una sola capa i la sortida depèn d'una combinació lineal de les entrades, o multicapa, amb múltiples capes incloent capes ocultes que processen les dades amb transformacions complexes, permetent l'aprenentatge de relacions no lineals entre dades.
A més, es diferencien per grau de connexió en xarxes totalment connectades, on totes les neurones d'una capa estan connectades amb les de la següent, i xarxes parcialment connectades, sense connexió total entre neurones de diferents capes.

[1] El sobreajustament (overfitting) es produeix quan un model predictiu s'ajusta massa als detalls i al soroll del conjunt de dades d'entrenament fins al punt de perdre la capacitat de generalitzar bé a noves dades.

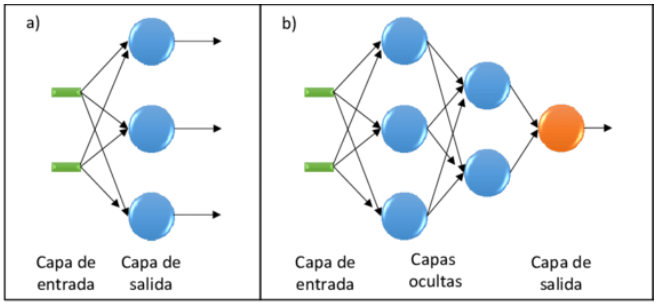


Figura 9. a) Xarxa neuronal monocapa, b) Xarxa neuronal multicapa

2. Objectius

L'objectiu principal d'aquest projecte és desenvolupar un model de predicció clínica, capaç de determinar la resposta al tractament SBRT en pacients amb càncer de pulmó. S'avaluaran diferents tècniques per escollir les millors variables amb les quals poder entrenar el model.

Més específicament, l'objectiu és realitzar una comparativa entre dos models predictius de complexitat diferent. El primer és una Xarxa Neuronal Convolucional (CNN), que és un model d'alta complexitat amb capacitats avançades de processament i anàlisi d'imatges. El segon és un model de Bosc Aleatori, que és relativament més senzill i es basa en múltiples arbres de decisió per fer prediccions. Aquesta comparativa permetrà determinar quin model és més eficaç per predir la resposta al tractament en pacients amb càncer de pulmó, utilitzant diferents tipus de dades clíniques i biomarcadors.

Addicionalment a la comparativa entre els dos models de predicció, l'estudi té com a objectiu examinar l'eficàcia dels models en funció de la diversitat de les dades amb les quals són entrenats. Això es realitzarà entrenant els models amb quatre conjunts diferents de dades, cada un incorporant diferents tipus d'informació biomèdica:

- Dades Clíniques solament: Utilitzant únicament les dades clíniques bàsiques dels pacients.
- Dades Clíniques i Radiòmiques: Combinant dades clíniques amb radiòmiques, que inclouen anàlisis detallades d'imatges mèdiques per extreure característiques que no són apreciables a simple vista.
- Dades Clíniques i Metabolòmiques: Integrant dades clíniques amb informació metabolòmica, que proporciona dades sobre els metabòlits presents en els teixits dels pacients, oferint una visió més profunda del seu estat metabòlic.
- Dades Clíniques, Radiòmiques i Metabolòmiques: Utilitzant una combinació de totes les dades disponibles, per veure si l'agregació d'aquestes informacions diverses millora la capacitat dels models per predir la resposta al tractament.

A través d'aquesta estratègia d'entrenament, es busca determinar quina combinació de tipus de dades proporciona els millors resultats predictius, i així optimitzar els models per a una aplicació clínica més eficaç i personalitzada.

3. Materials i mètodes.

Aquest projecte s'ha desenvolupat en diferents fases que són en primer lloc la definició del problema, en segon lloc la recol·lecció de dades, en tercer lloc el processament de les dades i finalment la creació i anàlisi del model en si.

En aquest cas, les respostes al tractament SBRT són quatre: resposta completa, parcial, sense resposta i empitjorament. S'entén la resposta completa com l'eliminació de qualsevol teixit tumoral en la zona tractada, la resposta parcial com una disminució del teixit tumoral en la zona tractada, sense resposta és que el tumor no ha crescut ni ha disminuït després del tractament i finalment, empitjorament vol dir que el teixit tumoral ha augmentat després del tractament. Aquestes quatre situacions s'han avaluat en tots els casos un mes i mig després de la radiació.

En aquest estudi es busca identificar els pacients que han tingut una resposta positiva i els que han tingut una resposta negativa, per això aquestes quatre opcions s'han agrupat en dues classes. Es defineix una resposta positiva al tractament si aquesta és completa o parcial i una resposta negativa si aquesta és sense resposta o empitjorament. D'aquesta manera, el model de classificació haurà de predir dos respostes únicament, si hi ha hagut resposta (0) o si no hi ha hagut resposta (1). La mostra consta de 65 pacients (n=65) tractats amb SBRT. De manera que un 71.4% han tingut una resposta al tractament (0) i un 28.6% que no han tingut resposta (1). Per tant, s'observa que es disposa d'una mostra desequilibrada amb una alta prevalença de zeros.

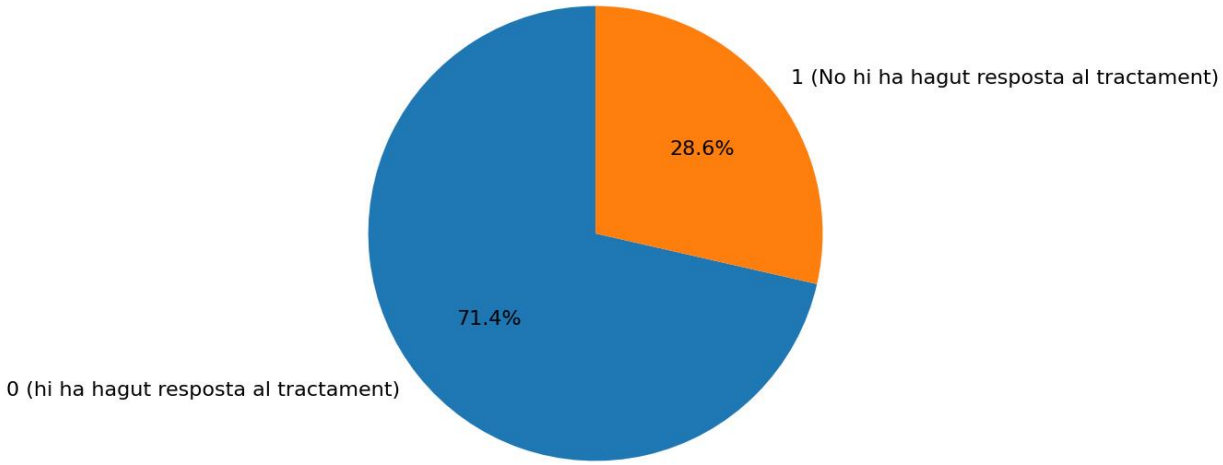


Figura 10. Distribució del tipus de resposta al tractament de la mostra.

3.1 Definició de les dades

Les dades s'obtenen a partir d'imatges TAC, en les que el facultatiu contorneja el volum del tumor (GTV) i el volum planificat per al tractament (PTV) (a), que inclou un marge geomètric. El PTV és utilitzat per dissenyar el tractament i impartir la dosi amb radioteràpia (b) SBRT. Els TACs i les estructures delineades s'han exportat en format DICOM i les característiques radiòmiques s'han extret mitjançant el programa 3DSlicer[2].

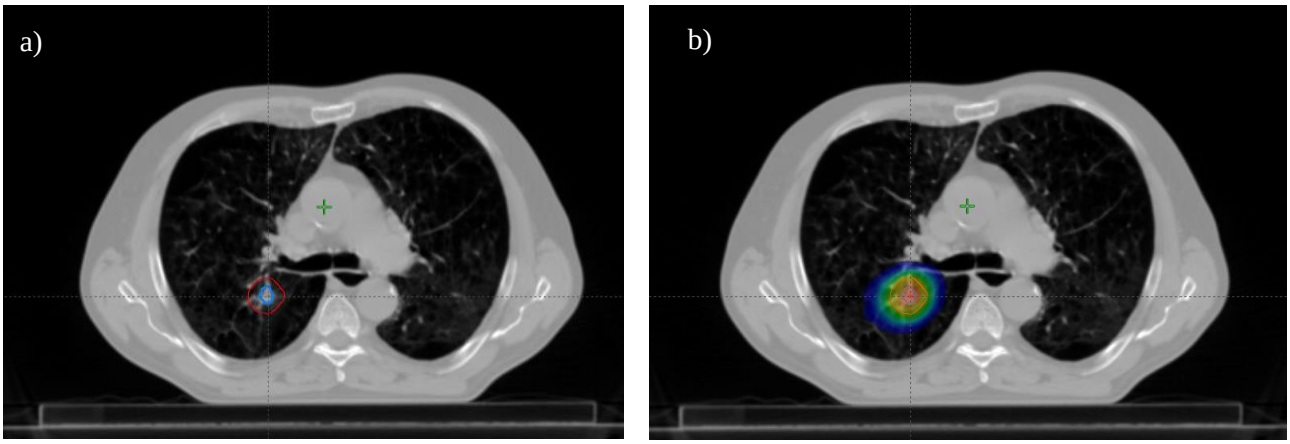


Figura 11. TAC axial amb PTV i GTV (a) i SBRT de pulmó amb dosi (b)

Les dades d'aquest projecte han estat anonimitzades seguint els estàndards legals a nivell del Reglament General de Protecció de Dades 2016/679 (RGPD) de la Unió Europea i la Llei Orgànica de Protecció de Dades Personals i garantia dels drets digitals 3/2018 (LOPDGDD) d'Espanya, així com la Llei d'Investigació Biomèdica 14/2007.

S'utilitzen tres tipus diferents de dades que es classifiquen en clíniques, radiòmiques i metabolòmiques. En total de 269 característiques. De les quals 15 són clíniques, 74 són dades metabolòmiques prèvies al tractament, 74 dades metabolòmiques posteriors al tractament i 107 són dades radiòmiques.

a) Dades clíniques: Aquestes dades s'extreuen de la història clínica del pacient.

- Edat
- Sexe
- Tabaquisme
- Enolisme
- DM
- HTA
- DLP
- Histologia
- Estadi clínic
- Estadi clínic general
- Localització
- Dosis total
- Fraccions
- PTV volum
- Immunoteràpia

[2] Programa de codi obert per a l'anàlisi i visualització d'imatges mèdiques, utilitzat per a l'extracció de característiques radiòmiques i la planificació de tractaments mèdics. Permet treballar amb dades DICOM i proporciona eines avançades per a la segmentació i modelatge 3D.

b) Dades radiòmiques: S'extreuen d'un VOI (*Volume of interest*), delimitat pel metge, que correspon al tumor. En concret al GTV *(Gross Tumor Volume*).

- *Elongation*
- *Flatness*
- *LeastAxisLength*
- *MajorAxisLength*
- *Maximum2DDiameterColumn*
- *Maximum2DDiameterRow*
- *Maximum2DDiameterSlice*
- *Maximum3DDiameter*
- *MeshVolume*
- *MinorAxisLength*
- *Sphericity*
- *SurfaceArea*
- *SurfaceVolumeRatio*
- *VoxelVolume*
- *GrayLevelVariance*
- *HighGrayLevelZoneEmphasis*
- *LargeAreaEmphasis*
- *LargeAreaHighGrayLevelEmphasis*
- *LargeAreaLowGrayLevelEmphasis*
- *LowGrayLevelZoneEmphasis*
- *SizeZoneNonUniformity*
- *SizeZoneNonUniformityNormalized*
- *10Percentile*
- *90Percentile*
- *Energy Entropy*
- *InterquartileRange*
- *Kurtosis*
- *Maximum*
- *MeanAbsoluteDeviation*
- *Mean Median Minimum*
- *Range*
- *RobustMeanAbsoluteDeviation*
- *RootMeanSquared*
- *Skewness*
- *TotalEnergy*
- *Uniformity*
- *Variance*
- *Autocorrelation*
- *ClusterProminence*
- *ClusterShade*
- *ClusterTendency*
- *Contrast*
- *Correlation*
- *DifferenceAverage*
- *DifferenceEntropy*
- *DifferenceVariance*
- *Id*
- *Idm*
- *LongRunHighGrayLevelEmphasis*
- *LongRunLowGrayLevelEmphasis*
- *LowGrayLevelRunEmphasis*
- *RunEntropy*
- *RunLengthNonUniformity*
- *RunLengthNonUniformityNormalized*
- *RunPercentage RunVariance*
- *ShortRunEmphasis*
- *ShortRunHighGrayLevelEmphasis*
- *ShortRunLowGrayLevelEmpha*
- *GrayLevelNonUniformity*
- *GrayLevelNonUniformityNormalized*
- *SmallDependenceEmphasis*
- *SmallDependenceHighGrayLevelEmphasis*
- *SmallDependenceLowGrayLevelEmphasis*
- *GrayLevelNonUniformity*
- *GrayLevelNonUniformityNormalized*
- *GrayLevelVariance*
- *HighGrayLevelRunEmphasis*
- *LongRunEmphasis*
- *SmallAreaEmphasis*
- *SmallAreaHighGrayLevelEmphasis*
- *SmallAreaLowGrayLevelEmphasis*
- *ZoneEntropy*
- *ZonePercentage*
- *ZoneVariance*
- *Busyness*
- *DependenceNonUniformityNormalized*
- *DependenceVariance*
- *GrayLevelNonUniformity*
- *GrayLevelVariance*
- *HighGrayLevelEmphasis*
- *LargeDependenceEmphasis*
- *LargeDependenceHighGrayLevelEmphasis*
- *LargeDependenceLowGrayLevelEmphasis*
- *LowGrayLevelEmphasis*
- *Imc2*
- *InverseVariance*
- *JointAverage*
- *JointEnergy*
- *JointEntropy*
- *MCC*
- *MaximumProbability*
- *Coarseness*
- *Complexity*
- *Contrast*
- *Strength*
- *SumAverage*
- *SumEntropy*

- *Idmn*
- *Idn*
- *Imc1*
- *SumSquares*
- *DependenceEntropy*

c) Metabolòmiques: Dades que aporten informació sobre metabòlits i s'extreuen a partir d'analítiques de sang.

- Pyruvic acid
- Lactic acid
- 2-Hydroxyisobutyric acid
- Glycolic acid Alanine
- 2-HydroxyButyric acid
- 3-methyl-2-oxobutyric acid (alphaketoisovaleric acid)
- 3-hydroxybutyric acid/3-hydroxyisobutyric acid
- 2-Hydroxyisovaleric acid
- 2-keto-3-methylvaleric acid
- 3-Hydroxyisovaleric acid
- Valine Benzoic acid
- Ethanolamine Leucine
- Phosphoric acid
- Glycerol
- Ethylmalonic acid
- Isoleucine
- Proline Glycine
- Succinic acid
- Glyceric acid
- Fumaric acid
- Serine Threonine
- Hydrocinnamic acid
- Malic acid
- d-Threitol
- Methionine
- Oxoproline
- 4-Hydroxyproline
- Threonic acid Erythronic acid
- DL-2-Hydroxyglutaric acid
- a-ketoglutaric acid
- Glutamic acid
- 4-Hydroxybenzoic acid
- Phenylalanine
- Dodecanoic acid
- d-Xylose
- Taurine
- d-Arabinose
- d-Xylitol
- d-Arabitol
- Glycerol-1-phosphate
- Glutamine
- 3-Phosphoglyceric acid
- Ornithine
- Citric acid
- Tetradecanoic acid
- Hippuric acid
- Vanillylmandelic acid
- 4-hydroxyPhenyllactic acid
- d-Fructose
- d-Mannitol
- d-Mannonic acid
- d-Galactitol
- Galacturonic acid
- Galactonic acid
- Saccharic acid
- Indole-3-propanoic acid
- Myo-Inositol
- Uric acid
- Sedoheptulose
- Indolelactic acid
- Linoleic acid
- Oleic acid
- Glucose 6-phosphate
- d-Sucrose
- Maltose
- a-Tocopherol
- Pyruvic acid
- Lactic acid
- 2-Hydroxyisobutyric acid
- Glycolic acid
- Alanine
- 2-HydroxyButyric acid
- 3-methyl-2-oxobutyric acid (alphaketoisovaleric acid)
- 3-hydroxybutyric acid/3-hydroxyisobutyric aci
- 2-Hydroxyisovaleric acid
- 2-keto-3-methylvaleric acid
- 3-Hydroxyisovaleric acid
- Valine Benzoic acid
- Ethanolamine

- Xylonic acid
- Ribonic acid
- Phosphoric acid
- Glycerol
- Ethylmalonic acid
- Isoleucine
- Proline Glycine
- Succinic acid
- Glyceric acid
- Fumaric acid
- Serine
- Threonine
- Hydrocinnamic acid
- Malic acid
- d-Threitol
- Methionine
- Oxoproline
- 4-Hydroxyproline
- Threonic acid
- Erythronic acid
- DL-2-Hydroxyglutaric acid
- a-ketoglutaric acid
- Glutamic acid
- 4-Hydroxybenzoic acid
- Phenylalanine
- Dodecanoic acid
- d-Xylose
- Taurine
- d-Arabinose
- d-Xylitol
- d-Arabitol
- Glycerol-1-phosphate
- Glutamine
- Xylonic acid
- Ribonic acid
- 3-Phosphoglyceric acid
- Ornithine
- Citric acid
- Tetradecanoic acid
- Hippuric acid
- Vanillylmandelic acid
- 4-hydroxyPhenyllactic acid
- d-Fructose
- d-Mannitol
- d-Mannonic acid
- d-Galactitol
- Galacturonic acid
- Galactonic acid
- Saccharic acid
- Indole-3-propanoic acid
- Myo-Inositol
- Leucine
- Sedoheptulose
- Indolelactic acid
- Linoleic acid
- Oleic acid
- Glucose 6-phosphate
- d-Sucrose
- Maltose
- Uric acid

3.2 Processat de dades.

El processat de dades és un pas molt important en la creació de models predictius. Aquest procés assegura la qualitat i eficàcia dels models i permet que aquest proporcioni conclusions fiables en situacions reals. En aquest projecte, el processat de dades s'ha dut a terme en dos fases. En la primera fase s'han eliminat manualment algunes característiques segons criteri clínic. Aquestes s'han considerat dades que no aporten informació rellevant, com per exemple la data de naixement del pacient.

Per el que fa a la segona fase, s'ha utilitzat la *LabelEncoder* de la llibreria *sklearn.preprocessing* de *Python* per codificar la variable clínica Estadi clínic general. Aquesta eina permet codificar etiquetes, d'aquesta manera es converteixen dades categòriques en nombres enters. Això facilita l'entrenament de l'algoritme predictiu, que en la majoria de casos esperen que l'entrada de dades sigui numèrica. De la mateixa llibreria *sklearn.preprocessing* utilitza la eina *StandardScaler* per escalar tots els valors de la matriu de dades. Aquesta funció s'utilitza per estandarditzar les dades de manera que tenen una mitja de zero i una desviació estàndard d'un. Aquesta normalització serveix per millorar el rendiment dels models de Machine Learning. Finalment, s'ha fet el tractament de dades buides. En aquest cas, degut al baix número de dades buides disponibles en el conjunt de dades, s'ha optat per eliminar-les.

La normalització, estandardització i codificació de les dades permeten que el funcionament dels models sigui més exacte, evitant *l'overfitting* i que el model generalitzi millor a noves dades.

3.3 Selecció de característiques

Els models predictius de Machine Learning aplicats a decisions clíniques solen tenir una gran quantitat de característiques. Identificar les variables més importants és un pas molt fonamental per reduir la complexitat de l'estudi, que es tradueix en menor necessitat de processat i menor temps d'entrenament. També s'evita el sobreajustament degut a soroll i es millora l'exactitud del model.

En aquest projecte s'ha fet un filtratge utilitzant quatre mètodes estadístics diferents. Aquest filtratge s'ha fet en base a l'estudi de Grissa, Dhouha, et al. "[13].

Aquest filtratge es basa en escollir característiques segons quatre mètodes estadístics i computacionals per identificar les característiques més rellevants. Cada un d'ells proporciona una llista, aquestes es comparen en una taula de presència-absència. Finalment, es seleccionen aquelles que han estat considerades importants per tres o quatre dels mètodes.

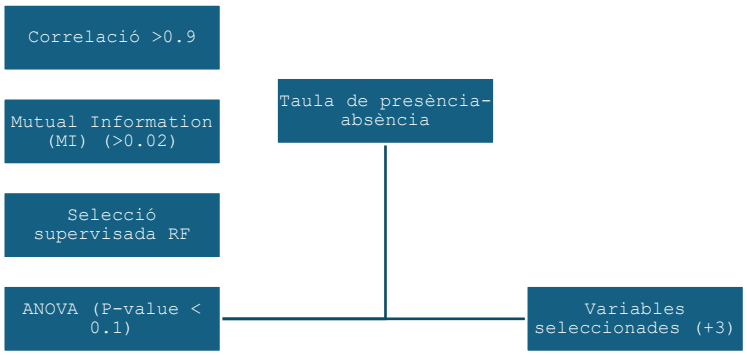


Figura 12. Esquema del procés de selecció de característiques.

3.3.1 Correlació

La correlació és una mesura estadística que ens indica com dues variables estan relacionades linealment. Es calcula utilitzant el coeficient de *Pearson*. El valor d'aquest coeficient pot variar entre -1 i 1, sent -1 una relació negativa perfecta i 1 una relació positiva perfecta. En el cas de 0, indica que no hi ha cap tipus de relació.

$$\rho X, Y = \frac{Cov(X,Y)}{\sigma X \sigma Y}$$

S'ha calculat aquest paràmetre amb la llibreria *numpy* de *Python*, s'han calculat totes les correlacions respecte la columna a predir i s'han escollit aquelles que presentaven un coeficient de *Pearson* (en valor absolut) major a 0.9.

3.3.2 Informació mútua (MI)

La informació mútua o MI és una mesura que indica la quantitat d'informació que una variable conté sobre una altra. Aquest mètode pot identificar relacions no lineals, a diferència del mètode de correlació, que només és capaç d'identificar relacions lineals. En aquest cas, si la resposta és 0, ens indica que les variables comparades són totalment independents.

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y)\, log(\frac{p(x,y)}{p(x)p(y)})$$

En aquest projecte s'ha utilitzat la eina *mutual_info_regression* de la llibreria *sklearn.feature_selection* de *Python*. Es seleccionen totes les variables que prenen un valor de MI més gran de 0.02 amb respecte a la variable a predir (y).

3.3.3 Selecció supervisada amb RF

Com s'ha explicat anteriorment en el projecte, els Random Forest o Boscos Aleatoris són un model d'aprenentatge automàtic basat en tasques de classificació i regressió. En aquest cas s'ha entrenat un model RF sobre el conjunt total de dades i amb la variables a predir com a objectiu. Aquest model proporciona una puntuació d'importància de cada característica, basada en com la inclusió de cada una millora la predicció del model. Generalment es mesura per la reducció mitjana d'impuresa. Les característiques que es seleccionen són les que més impacte tenen en la precisió del model. S'ha utilitzat la eina de *RandomForestClassifier* de la llibreria *sklearn.ensemble* de *Python*.

3.3.4 ANOVA

L'anàlisi de la variància (ANOVA) és una tècnica estadística utilitzada per determinar si hi ha diferències estadísticament significatives entre les mitjanes de tres o més grups independents. Aquest mètode treballa basant-se en dus hipòtesis: la nul·la, que postula que totes les mitjanes són iguals i la alternativa, que postula que que almenys una de les mitjanes dels grups és diferent. Aquestes hipòtesi s'avaluen amb els valors F-Statistic i P-Value. En aquest projecte, s'ha utilitzat únicament el valor P-Value, que indica si les diferències observades entre les mitjanes dels grups són estadísticament significatives.[14] Si el valor P és menor que el nivell de significança predeterminat (en aquest cas, 0.1), es rebutja la hipòtesi nul·la, suggerint que existeixen diferències significatives entre els grups estudiats. Per fer aquest càlcul s'ha utilitzat la eina *f_classif* de la llibreria *sklearn.features_selection* de *Python*.

3.3.5 Taula de presència-absència

Finalment, després d'aplicar els diferents mètodes estadístics descrits en els punts anteriors, els resultats s'organitzen en una taula de presència-absència. Aquesta taula serveix per visualitzar de manera efectiva quines característiques han estat identificades com a significatives segons cada mètode estadístic utilitzat. En aquesta taula, les columnes representen les diferents tècniques estadístiques aplicades, mentre que cada fila correspon a una característica específica del conjunt de dades. A la intersecció de cada fila i columna, es col·loca un '1' si la característica ha estat seleccionada com a important pel mètode corresponent, i un '0' si no ha estat seleccionada. Aquest format facilita la comparació ràpida entre els mètodes per determinar quines característiques són consistentment reconegudes com a rellevants a través dels diferents anàlisis.

	Correlation	Mutual Information	Random Forest	ANOVA
Pyruvic acid	1	1	1	0
Lactic acid	1	0	1	1
2-Hydroxyisobutyric acid	1	0	1	0
Glycolic acid	1	0	1	0
Alanine	1	1	1	0
2-HydroxyButyric acid	1	0	1	1
3-methyl-2-oxobutyric acid (alphaketoisovaleric acid)	1	1	1	1

Figura 13. Exemple de la taula presència absència del projecte.

A partir d'aquesta taula, s'han escollit totes les variables que han estat identificades com a important per a tres o quatre de les tècniques estadístiques.

3.4 Desenvolupament dels models predictius.

El desenvolupament dels models predictius en aquest estudi s'ha centrat en l'ús de tècniques d'aprenentatge profund per a la classificació binària de la resposta al tractament en pacients amb càncer de pulmó. S'han implementat dos models de diferent complexitat. Cal comentar que en ambdós casos s'ha utilitzat l'eina SMOTE (Synthetic Minority Over-sampling Technique) de la llibreria de imblearn.over_sampling de Python per crear grups de dades d'entrenament homogenis, degut a la gran disparitat entre resultats negatius i positius. Aquesta, funciona generant mostres sintètiques, en lloc de fer un sobremostrejat simple amb reemplaçament. Aquesta tècnica selecciona exemples que estan a prop en l'espai de característiques, calcula la diferència entre els exemples i multiplica aquesta diferència per un nombre aleatori que està entre 0 i 1. Després s'afegeix aquest producte a l'exemple minoritari per crear una nova mostra sintètica. Aquest procés ajuda a evitar el sobreajustament que pot ocórrer amb el sobremostrejat simple i fa que el model sigui més robust en predir les dades noves, mantenint al mateix temps una bona distribució i variabilitat dins de la classe minoritària.

3.4.1 Model de xarxa neuronal.

S'ha creat un model basat en xarxes neuronals denses utilitzant la biblioteca *Keras* de *Python*, un entorn d'alt nivell per a xarxes neuronals convolucional 1D (Conv1D) construïda amb l'API seqüencial de *Keras*, dissenyada principalment per a gestionar dades seqüencials. Comença amb una capa Conv1D que té 64 filtres i una mida de *kernel* de 2, utilitzant la funció d'activació *ReLU* per introduir no linealitats. Aquesta capa és seguida per una capa MaxPooling1D amb una mida de pool de 2 per a reduir la dimensionalitat i prevenir el sobreajustament. La seqüència de dades processada s'aplana en un vector unidimensional per la capa *Flatten*, que després passa a través d'una capa densa de 50 unitats amb activació *ReLU* per a una major transformació no lineal. Finalment, el model conclou amb una capa de sortida densa d'una sola unitat, típicament utilitzada per a tasques de classificació en dades seqüencials.

Per a millorar la convergència durant l'entrenament, s'ha utilitzat l'optimitzador *Adam*. Aquest optimitzador permet millorar la eficiència en tasques d'aprenentatge profund. El model s'ha compilat utilitzant la funció de pèrdua 'binary_crossentropy', que és essencial per a tasques de classificació binària. La funció d'entropia creuada binària mesura la distància entre les probabilitats predites i els valors reals (0 o 1 en aquest cas), penalitzant les prediccions incorrectes que s'allunyen de la realitat esperada. El propòsit d'utilitzar aquesta funció de pèrdua és minimitzar l'error en les prediccions, ajustant els pesos de manera que el model assigni la major probabilitat a la classe correcta. Això ajuda a garantir que el model sigui eficient al distingir entre respostes positives i negatives al tractament.

A més, s'ha monitoritzat la 'balanced_accuracy' com a mètrica de rendiment per a garantir que el model sigui just amb ambdues classes. Aquesta, s'ha definit utilitzant la biblioteca *backend* de *Keras* per calcular la sensibilitat i especificitat, obtenint així un mesurament més equilibrat del rendiment en classificacions binàries. S'ha implementat de la següent manera:

```
from keras import backend as K

def balanced_accuracy(y_true, y_pred):

    # Umbraliza las predicciones para obtener la clase binaria predicha
(0 o 1)
    threshold = 0.5
    y_pred_thresholded = K.cast(K.greater(y_pred, threshold), K.floatx())

    # Calcula True Positives, False Positives, True Negatives, y False
Negatives
    tp = K.sum(y_true * y_pred_thresholded)
    tn = K.sum((1 - y_true) * (1 - y_pred_thresholded))
    fp = K.sum((1 - y_true) * y_pred_thresholded)
    fn = K.sum(y_true * (1 - y_pred_thresholded))

    # Calcula Sensitivity (Recall) y Specificity
    sensitivity = tp / (tp + fn + K.epsilon())
    specificity = tn / (tn + fp + K.epsilon())
    balanced_acc = (sensitivity + specificity) / 2
    return balanced_acc
```

Codi 1. Codi funció balanced accuracy.

També s'ha incorporat un mecanisme de parada anticipada, mitjançant l'ús de *EarlyStopping* de *Keras* en forma de *callback*[3]. Aquest mecanisme monitoritza la pèrdua de validació i interromp l'entrenament si no es detecten millores substancials durant 10 èpoques consecutives, evitant així el sobreajustament i garantint que el model conservi els pesos que hagin obtingut el millor rendiment en el conjunt de validació.

El model s'ha entrenat utilitzant una estratègia de validació creuada amb *KFold*, dividint les dades en 5 subconjunts per garantir que l'avaluació del model sigui robusta i menys susceptible a

[3] Funció que es crida a determinats punts durant el cicle de vida de l'entrenament d'un model, com al final de cada època. Permet realitzar accions específiques o comprovar l'estat del model en moments crítics.

variacions específiques del conjunt de dades. Això implica entrenar el model múltiples vegades, cada cop amb un subconjunt diferent de dades com a conjunt de prova i els altres com a entrenament. Aquest procediment és útil quan les dades que es disponibles són desequilibrades i/o escasses, com el cas que escau en aquest treball.

3.4.2 Model Random Forest

S'ha definit també un model predictiu basat en *Random Forest*. Per a l'entrenament del model, les dades han estat dividides en conjunts d'entrenament i validació utilitzant la funció *train_test_split* de la biblioteca *sklearn.model_selection.* Aquesta funció segmenta les dades en proporcions especificades, en aquest cas, un 80% per a entrenament i un 20% per a validació. La segmentació és crucial per validar la capacitat del model de generalitzar a noves dades que no ha vist durant l'entrenament.

S'ha configurat el model *RandomForestClassifier* amb 100 arbres de decisió, valor típicament utilitzat en aplicacions mèdiques, proporcionant una bona equilibri entre rendiment de còmput i capacitat predictiva. El paràmetre *random_state* fixat a 42 assegura que els resultats siguin repetibles, facilitant la consistència en les execucions múltiples del model. Després de la configuració, el model s'entrena amb les dades d'entrenament, ajustant els arbres de decisió per optimitzar la classificació.

S'ha utilitzat l'atribut *feature_importance* d'un model *RandomForestClassifier* que proporciona una mesura de la importància de cada característica utilitzada per les decisions de l'arbre durant el procés de predicció.

3.5 Validació dels models.

Aquest treball ha examinat dos models predictius diferents utilitzant diverses combinacions de dades clíniques, metabolòmiques i radiòmiques per a predir la resposta a la teràpia contra el càncer a curt termini. Els models s'utilitzen diferents mètriques de validació, entre les que es troben la sensibilitat i la especificitat. S'ha definit com a resposta 0 aquells que tindran una resposta completa o parcial al tractament i coma resposta 1 a aquells que tindran una resposta estable i progressió de la malaltia. S'entén la sensibilitat com la taxa de veritable positiu. És a dir, en aquest cas, mesura la capacitat per identificar correctament els individus que tenen una resposta negativa al tractament, en aquest cas classificats com a 1:

$$Sensibilitat = \frac{Veritables\ Positius}{Veritables\ Positius + Falsos\ Negatius}$$

S'entén la especificitat, en aquest cas, com la capacitat del model per detectar correctament els casos el càncer ha respost efectivament al tractament (0).

$$Especificitat = \frac{Veritables\ Negatius}{Veritables\ Negatius + Falsos\ Positius}$$

Per aquesta aplicació interessa optimitzar la sensibilitat, ja que el cost clínic de no detectar correctament una "no resposta" al tractament (progressió de la malaltia) és alt. Una predicció de "no resposta" permetrà valorar altres tractaments o intervencions que millorin la resposta final.

A més, es realitza una anàlisi de les matrius de confusió per cada model. Aquestes matrius proporcionen una vista detallada de la relació entre les etiquetes predites i les reals, incloent els veritables positius (TP) , falsos positius (FP), veritables negatius (TN) i falsos negatius (FN). Aquesta informació és vital per comprendre en quines condicions específiques un model pot fallar i permet

ajustar els paràmetres o estratègies de modelatge per millorar el rendiment. En el cas del model de xarxa neuronal, aquesta matriu de confusió esta formada per la mitjana de les matrius obtingudes en cada iteració del model.

	Predicció Positiva (1)	**Predicció Negativa (0)**
Condició Positiva (1)	Verdader Positiu (TP)	Fals Negatiu (FN)
Condició Negativa (0)	Fals Positiu (FP)	Verdader Negatiu (TN)

Figura 14. Exemple de matriu de confusió.

4. Resultats

Un total de 65 pacients amb càncer de pulmó van ser tractats amb SBRT. Dels quals un 71.4% han tingut una resposta al tractament i un 28.6% no han tingut resposta al tractament. S'han recollit un total de 270 característiques. De les quals 15 són clíniques, 74 són dades metabolòmiques prèvies al tractament, 74 dades metabolòmiques posteriors al tractament i 107 són dades radiòmiques. Aquestes s'han dividit en quatre grups diferents. El primer que consta de les 15 característiques clíniques únicament, el segon amb 122 característiques entre les que es troben les clíniques i les radiòmiques, el tercer amb163 característiques en les que hi ha les clíniques i les metabolòmiques i finalment un amb les 270 que consta de les clíniques, metabolòmiques i radiòmiques.

Les variables seleccionades a partir del mètode descrit anteriorment sobre el *dataset* de dades clíniques són 5, que corresponen 33,33% del total. Aquestes són: *'HTA', 'DLP', 'HISTOLOGIA', 'LOCALIZACIÓN ', 'FRACCIONES'*. S'ha calculat la correlació d'aquestes amb la resposta, com es pot veure en la figura 15.

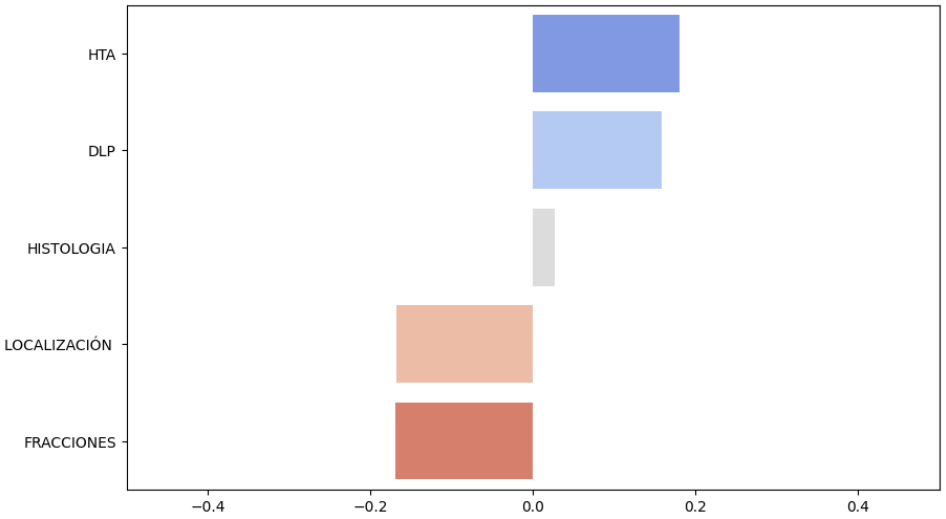


Figura 15. Correlació de les dades seleccionades a partir de dades clíniques.

Les variables seleccionades a partir del mètode descrit anteriorment sobre el *dataset* de dades clíniques i radiòmiques són 25, que corresponen al 20.5% del total:

a) Clíniques:

'INMUNOTERAPIA'

b) Radiòmiques:

'LeastAxisLength', 'MajorAxisLength', 'Maximum2DDiameterColumn', 'Maximum2DDiameterRow', 'Maximum2DDiameterSlice', 'Maximum3DDiameter', 'MeshVolume', 'MinorAxisLength', 'SurfaceVolumeRatio', 'VoxelVolume', '10Percentile', 'Minimum', 'Range', 'Imc1', 'DependenceNonUniformityNormalized', 'GrayLevelNonUniformity', 'SmallDependenceEmphasis', 'GrayLevelNonUniformity.1', 'RunLengthNonUniformityNormalized', 'GrayLevelNonUniformity.2', 'SizeZoneNonUniformityNormalized', 'ZoneVariance', 'Coarseness', 'Contrast.1'.

S'ha calculat la correlació d'aquestes amb la resposta, com es pot veure en la figura 16.

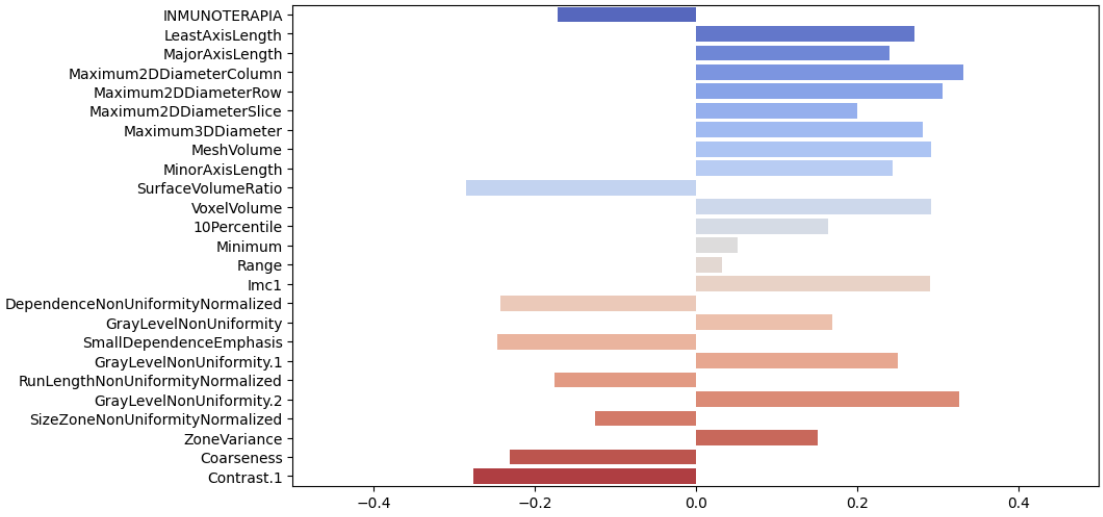


Figura 16. Correlació de les dades seleccionades a partir del conjunt de dades clíniques i radiòmiques amb la variable a predir pel model.

Les variables seleccionades a partir del mètode descrit anteriorment sobre el *dataset* de dades clíniques i metabolòmiques són 10, que corresponen al 5.89% del total:

a) Clíniques:

'FRACCIONES', 'INMUNOTERAPIA',

b) Metabolòmiques:

'Pyruvic acid ', 'Benzoic acid ', 'Glycerol-1-phosphate ', 'Glutamine ', 'Galacturonic acid ', '3-Hydroxyisovaleric acid .1', 'Oxoproline .1', 'Glutamine .1'. S'ha calculat la correlació d'aquestes amb la resposta, com es pot veure en la figura 16.

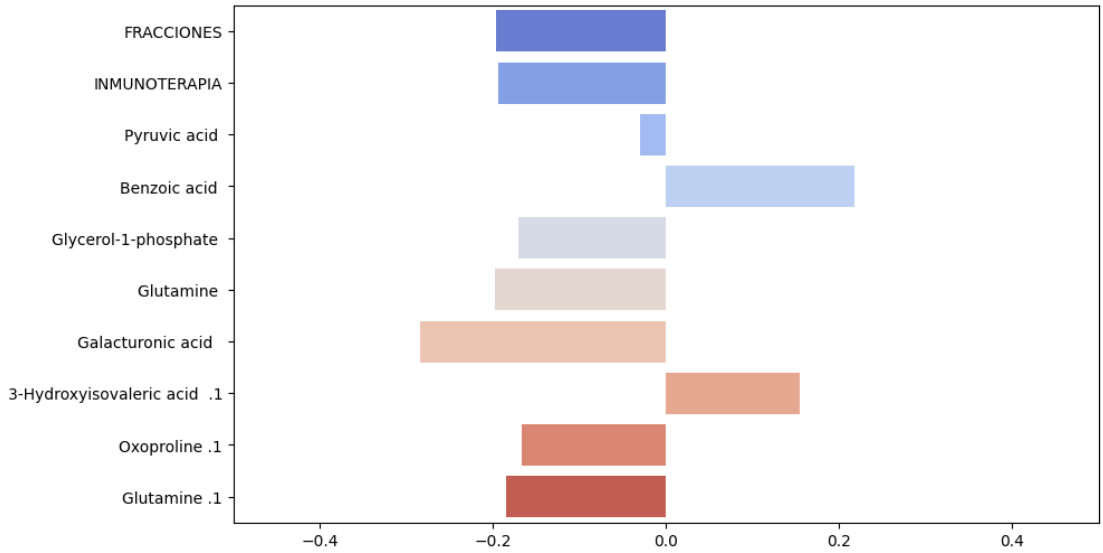


Figura 17. Correlació de les dades seleccionades a partir del conjunt de dades clíniques i metabolòmiques amb la variable a predir pel model.

Les variables seleccionades a partir del mètode descrit anteriorment sobre el *dataset* de dades clíniques, radiòmiques i metabolòmiques són 28, que corresponen al 10.37% del total:

a) Clíniques:

FRACCIONES, INMUNOTERAPIA

b) Metabolòmiques:

'Benzoic acid ', 'Glycine ', 'Threonic acid ', 'Dodecanoic acid ', 'Sedoheptulose ', 'Linoleic acid ', 'd-Threitol .1', 'Oxoproline .1', 'd-Arabitol .1', 'Glutamine .1', 'Ornithine .1', 'Uric acid .1',

c) Radiòmiques:

'Elongation', 'Flatness', 'LeastAxisLength', 'MajorAxisLength', 'Maximum2DDiameterColumn', 'Maximum2DDiameterSlice', 'Sphericity', 'GrayLevelNonUniformity', 'LongRunEmphasis', 'ShortRunEmphasis', 'LargeAreaEmphasis', 'SizeZoneNonUniformity', 'ZoneVariance', 'Contrast.1'. S'ha calculat la correlació d'aquestes amb la resposta, com es pot veure en la figura 17.

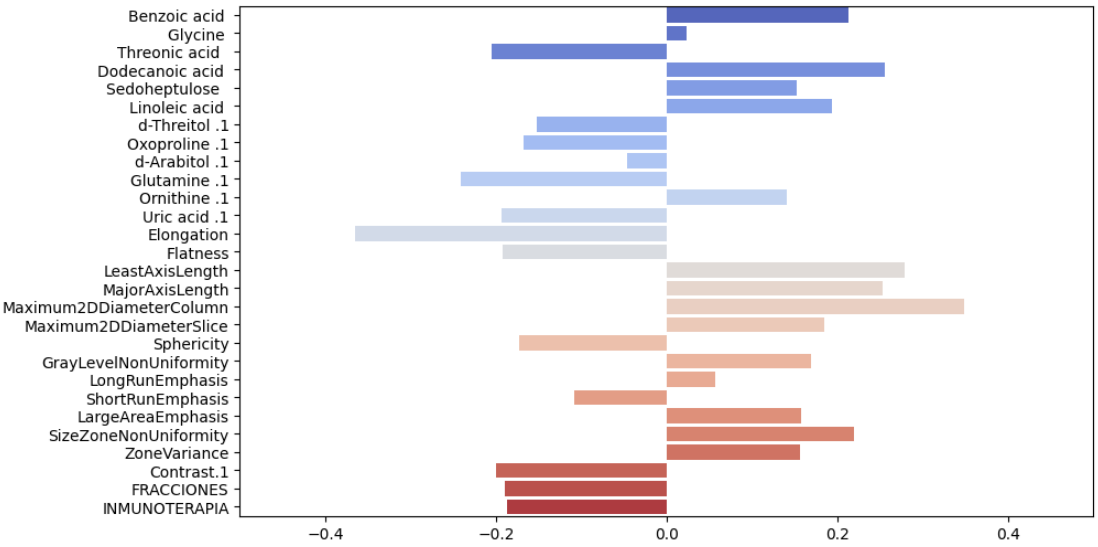


Figura 18. Correlació de les dades seleccionades a partir del conjunt de dades clíniques, metabolòmiques i radiòmiques amb la variable a predir pel model.

Un cop escollides les variables s 'han entrenat i testejat dos models de predicció diferents. Per un costat, una CNN amb la llibreria *keras* de *Python*. En la que s'ha iterat varies vegades en les dades per poder entrenar i testejar el model en diferents iteracions. A més, d'optimitzar-lo mitjançant el càlcul de *Balanced Accuracy*[4] (BA) i utilitzant la tècnica d'*early stopping*, de manera que si la BA no millorava en 10 capes del model, aquest deté l'entrenament i així evitar el sobreajustament d'aquest. Per altra banda, s'ha definit un model d'arbre de decisió amb la llibreria *sklearn* de *python*. Els resultats es resumeixen en la següent taula.

[4] El Balanced Accuracy és una mètrica que calcula la Mitjana de les taxes de veritables positius i negatius per a cada classe. D'aquesta manera ajusta el rendiment del model en conjunts de dades poc balancejades.

	CNN - Keras		Random Forest - Sklearn	
	Especificitat	**Sensibilitat**	**Especificitat**	**Sensibilitat**
Clíniques	1	0.44	0.89	0.67
Clíniques i radiòmiques	0.56	0.50	0.75	0
Clíniques i metabolòmiques	0.60	0.75	0.75	0.50
Cliques, metabolòmiques i radiòmiques	0.12	1	0.78	1

Taula 1. Comportament dels dos models treballats segons especificitat i sensitivitat.

De la mateixa manera, s'ha buscat la matriu de confusió de cada model entrenat amb diferents conjunts de dades. S'ha obtingut amb les llibreries *confusion_matrix* i *ConfusionMatrixDisplay* de la eina *sklear.metrics* de *Python*.

a) Clíniques.

En aquest cas, la matriu del model CNN (figura 19), indica una bona capacitat per predir la classe 0 (4 TN) però en el cas de la classe 1 detecta 2 FP i 5 TP. La segona matriu, del model RF (figura 20), mostra un millor equilibri, amb 8 TN per la classe 0 i 1 TP per la classe 1, però amb algunes prediccions incorrectes (1 FN i 2 FP).

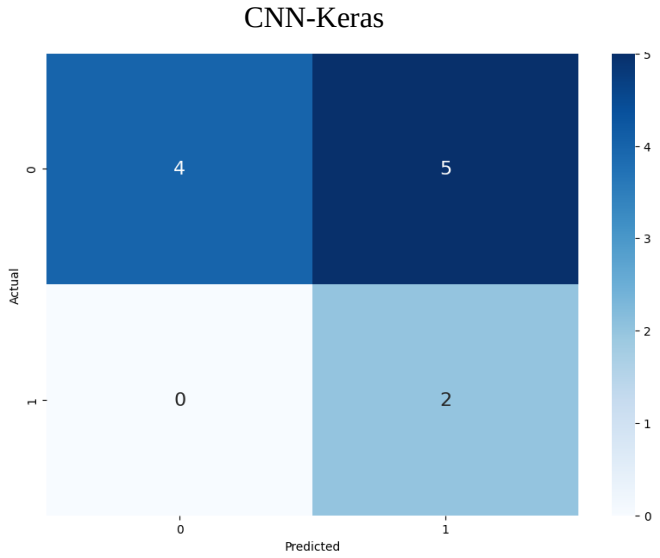


Figura 19. Matriu de confusió del model CNN amb dades clíniques.

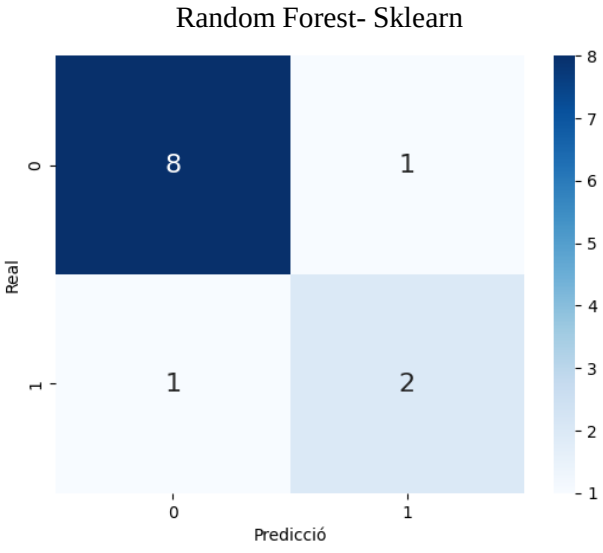


Figura 20. Matriu de confusió del model Random Forest amb dades clíniques.

b) Clíniques i radiòmiques.

La matriu del model CNN (figura 21) destaca en predir la classe 0 amb 5 TN, també la classe 1 amb 4 TP, però es troba 1 FN i 1 FP respectivament. La matriu del model *Random Forest* (figura 22) amb 6 TN per la classe 0 i 2 TP per la classe 1, però també mostra 3 FP, mentre que cap FN. Ambdues matrius indiquen problemes per detectar correctament la classe 1, amb la segona mostrant una millora lleu en aquest aspecte.

CNN-Keras

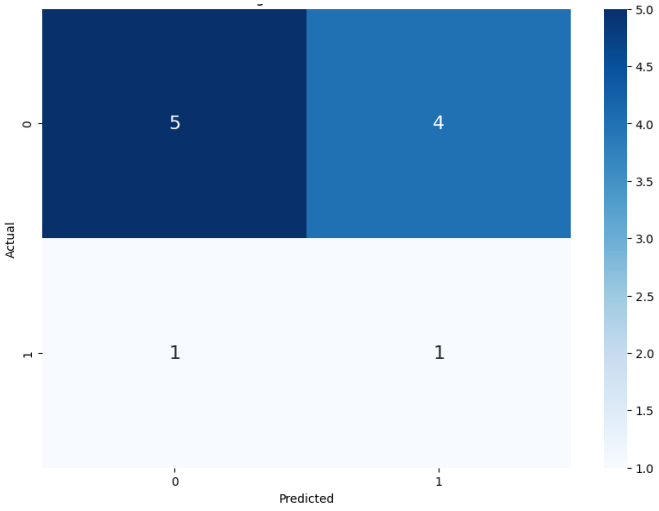


Figura 21. Matriu de confusió del model CNN amb dades clíniques i radiòmiques.

Random Forest-Sklearn

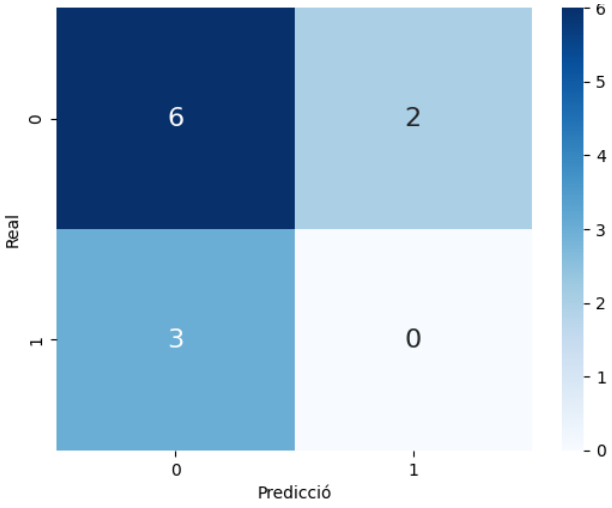


Figura 22. Matriu de confusió del model Random Forest amb dades clíniques i radiòmiques..

c) Clíniques i metabolòmiques.

La matriu del model CNN (figura 23), mostra una forta capacitat de predicció per la classe 0 amb 3 TN, però té dificultats amb la classe 1 amb 1 FN,i 3 FP, identifica correctament 2 vegades la classe 1. La matriu del model *Random Forest* (figura 24), mostra un major equilibri, amb 6 TN per la classe 0 i 2 TP per la classe 1, però també presenta errors amb 1 FP i 1 FN. Aquestes matrius reflecteixen enfoques diferents: la primera prioritzant la classe 0 i la segona buscant un equilibri millor entre les dues classes.

CNN – Keras

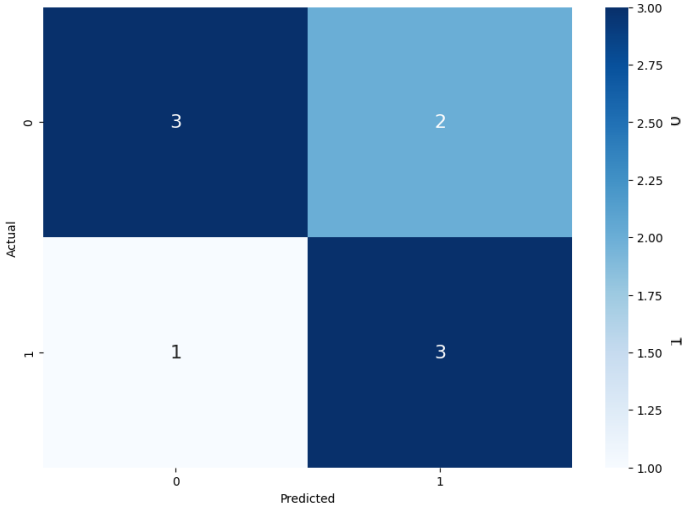


Figura 23. Matriu de confusió del model CNN amb dades clíniques i metabolòmiques.

Random Forest-Sklearn

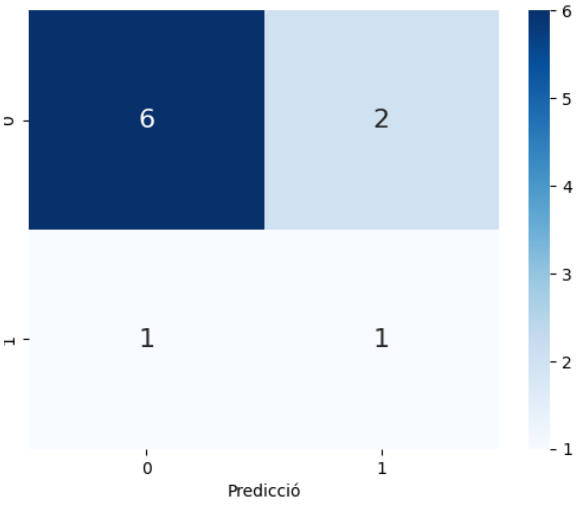


Figura 24. Matriu de confusió del model Random Forest amb dades clíniques i metabolòmiques.

d) Clíniques, metabolòmiques i radiòmiques.

La matriu del model CNN (figura 25), mostra una alta eficàcia per detectar la classe 0 amb 1 TN, sense errors (0 FN) en la predicció de la classe 1, però amb una presència de 7 FP que indiquen confusió en algunes prediccions. La matriu del *Random Forest* (figura 26), indica que el model també és efectiu en detectar la classe 0 amb 7 TN, però presenta 1 FP, mostrant una certa confusió al predir la classe 1 com 0. A més, detecta correctament la classe 1 en 2 cas (TP), millorant respecte a la primera matriu.

CNN-Keras

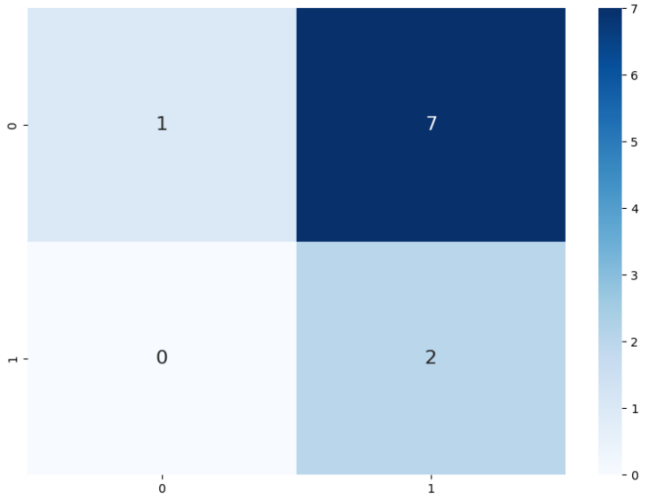


Figura 25. Matriu de confusió del model CNN amb dades clíniques, metabolòmiques i radiòmiques.

Random Forest-Sklearn

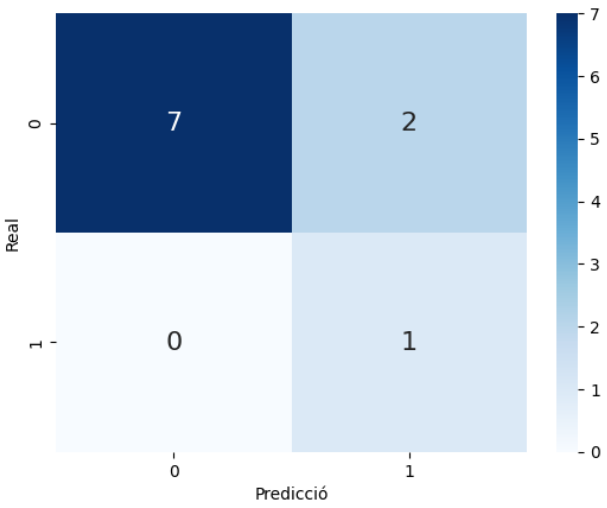


Figura 26. Matriu de confusió del model Random Forest amb dades clíniques, metabolòmiques i radiòmiques.

Les matrius de confusió del model CNN, tot i utilitzar *datasets* d'entrenament equilibrats, reflecteixen una capacitat limitada dels models per predir correctament la classe 1. Majoritàriament, mostren un nombre significatiu de Falsos Positius (FP), on la classe 1 és incorrectament predita com a classe 0. Per aquest motiu s'ha estudiat l'evolució del paràmetre *Balanced Accuarcy* en les diferents iteracions del model (Annex II).

S'observa en ambdós models, que el millor funcionament es dona quan el model s'entrena amb els tres tipus de dades. A més, el model amb un millor compromís entre especificitat i sensitivitat el proporciona el model construït a través d'un arbre de decisió. Les importàncies de cada característica en aquest model, quan s'utilitzen els tres tipus de dades estan representades a la figura 26. S'observa com en les 5 característiques més importants, es troben tant variables metabolòmiques com radiòmiques.

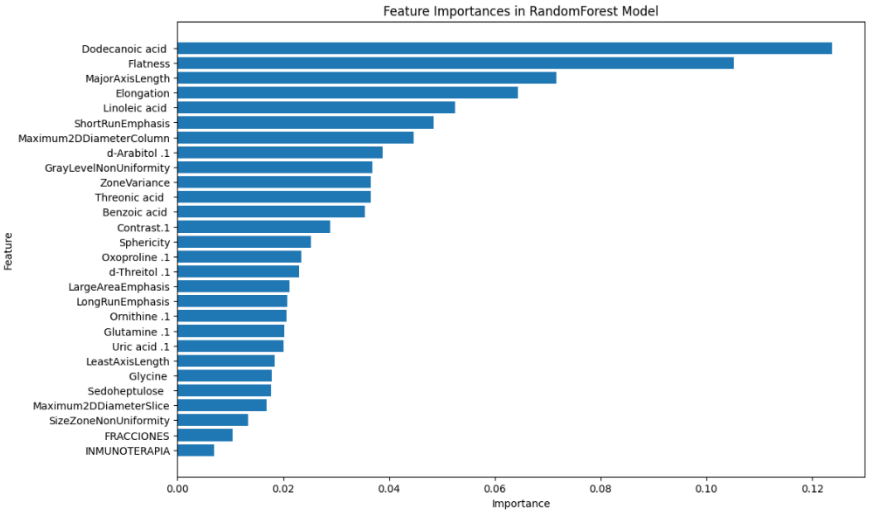


Figura 27. Importància de les característiques en la presa de decisions del model Random Forest.

.

5. Discussió

Es proposa aquest treball per donar resposta a un problema de decisió clínica en pacients amb càncer de pulmó. Existeix una necessitat a nivell assistencial per poder anticipar-se a com reaccionarà cada pacient al tractament amb SBRT. El poder identificar de forma precoç aquest comportament proporcionaria als clínics el poder generar un pla terapèutic personalitzat, a part de valorar altres vies de tractament i evitar els efectes secundaris que suposa aquest tractament radiològic. D'altra banda, aquesta eina podria disminuir les carregues de feina dels professionals de la salut agilitzant i donant suport als processos de decisió clínica. El propòsit d'això és millorar l'eficiència dels tractaments i donar una atenció personalitzada i segura a un nombre més elevat de pacients.

Cal destacar que el model de selecció de característiques ha demostrat una capacitat consistent per identificar variables rellevants de tots els tipus amb els quals ha estat alimentat. Així, en situacions on s'han utilitzat variables clíniques i radiòmiques, el model ha reconegut com a importants tant variables clíniques com radiòmiques. Aquest patró s'ha repetit en altres conjunts de dades, seleccionant característiques pertinents tant de les dades clíniques com de les metabolòmiques i altres tipus de dades disponibles, garantint una representació equilibrada de totes les dimensions analitzades.

S'ha observat una gran disparitat en la resposta al tractament. Això significa que s'ha treballat amb un conjunt de dades poc balancejades, en el que s'ha hagut d'utilitzar mètodes per pal·liar aquest efecte i poder entrenar els models de manera homogènia en tots els casos. Tot i haver utilitzat aquestes tècniques, s'observa una dificultat clara en tots els casos per predir els casos positius (resposta 1). Que en aquesta aplicació corresponen a aquells que no han tingut cap resposta al tractament. El que comporta un risc en la pràctica clínica. Per mitigar aquest problema es poden utilitzar tècniques com el *data warping*[5] o la generació de dades sintètiques. Aquestes estratègies podrien ajudar a augmentar artificialment la grandària de la mostra, proporcionant al model una millor oportunitat per aprendre les característiques discriminatives de totes les classes. No obstant això, aquestes tècniques també poden portar a la creació de dades que no reflecteixen acuradament la realitat clínica, el que podria resultar en models que funcionen bé en dades de test però que fallen en l'entorn clínic real.

Una altra via a explorar és la realització d'estudis multicèntrics que no només aportarien un augment en el volum de les dades disponibles per a l'anàlisi sinó que també milloraria la variabilitat i la representativitat de les dades.

S'ha vist que el model de xarxes neuronals convolucionals no aconsegueix aprendre correctament de les dades, sobretot en el cas en que la resposta és 1, en què en la majoria d'ocasions es reconeixen 0 TP. Això, pot estar relacionat en que les CNN requereixen grans quantitats de dades etiquetades per entrenar-se efectivament, i la mancança d'un volum suficient pot impedir que el model aprengui les característiques essencials necessàries per a una bona generalització. Aquesta situació es complica encara més quan el conjunt de dades presenta un gran desequilibri en les classes, és a dir, quan una de les classes està molt més representada que l'altra. Aquest mal comportament es podria relacionar amb el fet de que els models CNN son més eficients en imatges.

Potser per aquests motius el model RF dona una millor resposta en aquesta aplicació, degut tant a la menor dependència de la quantitat de les dades d'aquest tipus de tècniques com a la seva millor resistència al sobreajustament.

Tot i que el model de Xarxa neuronal, no hagi tingut el resultat esperat, sí que s'observa una millora dels resultats a mesura que s'utilitzen més tipus de dades. El mateix s'observa en el model

[5] Tècnica de processament de dades que implica la transformació no lineal de les dades existents per augmentar la diversitat dins d'un conjunt de dades d'entrenament.

de *RandomForest*, que quan s'entrena amb tot tipus de dades dona un resultat de sensibilitat òptim. És a dir, el model és capaç de reconèixer la resposta 1 (sense resposta al tractament) en el 100% dels casos. A més, en aquesta opció, s'observa que en les 5 variables més importants en la presa de decisions del model, s'hi troben tant dades metabolòmiques com radiòmiques. El que reforça la hipòtesi inicial de la importància de treballar amb els tres tipus de dades combinades.

Tot i les limitacions comentades, aquest estudi confirma la utilitat dels models avançats de dades com a eines per als sistemes de decisió clínica. Els resultats destaquen la importància d'incorporar dades de diferents tipus, com les dades radiòmiques i metabolòmiques, que poden enriquir els models i proporcionar perspectives més profundes sobre la malaltia, que les dades clíniques per si soles poden no ser capaces de revelar. Aquestes fonts de dades diverses poden millorar significativament la precisió dels models predictius i, per tant, proporcionar una millor eina de suport en la presa de decisions clíniques.

És essencial continuar amb la investigació i desenvolupament en aquesta àrea. Els estudis futurs haurien d'explorar com integrar de manera més efectiva aquestes quan es segueixen estratègies multi-òmiques, com superar els reptes tècnics i ètics associats amb la seva implementació en la pràctica clínica diària. La col·laboració interdisciplinària serà clau per avançar en aquest camp i per assegurar que els avenços en la intel·ligència artificial i l'anàlisi de dades tinguin un impacte positiu en el tractament del càncer.

6. Conclusions

En conclusió, s'ha vist que els models CNN poden trobar-se amb problemes quan les dades amb les que treballen no estan ben equilibrades o quan el tamany de la mostra no és suficient. Això posa de manifest la necessitat d'usar tècniques que ajudin a fer que les dades siguin més completes i representatives. També s'observa que en alguns casos, els models RF poden resultar més efectius, ja que treballen millor quan existeixen desequilibris de les classes i poden funcionar millor amb tamanys de mostra menors. L'estudi també recalca la importància de continuar investigant com es poden integrar diferents tipus de dades, com les radiòmiques i les metabolòmiques, per fer que els models predictius siguin més rics i personalitzats. Els millors resultats en ambdós models es donen quan es treballen amb els tres tipus de dades, fet que indica la convivència d'apostar per tècniques multi-òmiques.

Finalment, queda clar que per aprofitar al màxim els models de CNN i altres tècniques de modelatge en medicina, necessitem més treball conjunt entre investigadors, metges i tecnòlegs. Avançar en aquest camp requerirà crear conjunts de dades més grans i equilibrats i explorar noves maneres d'analitzar les dades. Treballar de manera multidisciplinària serà essencial per convertir els avenços tecnològics en avantatges reals per als pacients. En concret, els models predictius basats en intel·ligència artificial poden ser de gran utilitat coma eines de suport a les decisions clíniques.

Referencies

[1] Nasser, Ibrahim M., and Samy S. Abu-Naser. "Lung cancer detection using artificial neural network." *International Journal of Engineering and Information Systems (IJEAIS)* 3.3 (2019): 17-23.

[2] Hamada A. A. Noreldeen, Xinyu Liu, Guowang Xu. “Metabolomics of lung cancer: Analytical platforms and their applications”. JOURNAL OF SEPARATION SCIENCE (2020): 120-135

[3] Collins, Lauren G., et al. "Lung cancer: diagnosis and management." American family physician 75.1 (2007): 56-63.

[4] Sher, Taimur, Grace K. Dy, and Alex A. Adjei. "Small cell lung cancer." Mayo Clinic Proceedings. Vol. 83. No. 3. Elsevier, 2008.

[5] American Cancer Society, "What Is Lung Cancer?” American Cancer society, 2024.

[6] Shinde, Ashwin, et al. "Stereotactic body radiation therapy (SBRT) for early-stage lung cancer in the elderly." Seminars in oncology. Vol. 45. No. 4. WB Saunders, 2018.

[7] Thompson, Marcher, and Kenneth E. Rosenzweig. "The evolving toxicity profile of SBRT for lung cancer." Translational lung cancer research 8.1 (2019): 48.

[8] Méndez-Rodríguez, Karen Beatriz, et al. "Metabolómica como nueva herramienta para el diagnóstico oportuno en enfermedades no transmisibles." Revista de Salud Ambiental 19.2 (2019): 109-115.

[9] Rivas, Solange, and Ricardo Armisén. "El cáncer de pulmón de células no pequeñas en la era de la medicina de precisión." Revista Médica Clínica Las Condes 33.1 (2022): 25-35.

[10] Martí-Bonmatí, Luis, coordinador. *RADIÓMICA*. Fundación Instituto Roche, 2022, www.institutoroche.es. Informe Anticipando.

[11] Refaee, Turkey, et al. "El papel emergente de la radiómica en la EPOC y el cáncer de pulmón." *Kompass Neumología* 2.2-3 (2020): 46-53.

[12] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[13] Grissa, Dhouha, et al. "Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data." Frontiers in molecular biosciences 3 (2016): 30.

[14] Box, G. E., Hunter, W. H., & Hunter, S. (1978). Statistics for experimenters (Vol. 664). New York: John Wiley and sons.

Annex I

En aquest annex s'adjunten les taules de presència absència obtingudes en el model de selecció de característiques.

a) Clíniques

	Correlational_corr	Mutual Information	Random Forest	ANOVA
ID	0	0	1	0
RESPUESTA POR TC A 1.5 MESES	1	0	1	0
EDAD	0	0	0	0
SEXO	0	0	0	0
TABAQUISMO	0	1	0	0
ENOLISMO	0	0	0	0
DM	0	0	0	0
HTA	1	1	0	0
DLP	1	1	1	0
HISTOLOGIA	0	1	1	0
ESTADIAJE CLÍNICO (T)	0	0	0	0
ESTADI CLÍNIC general	0	0	0	0
LOCALIZACIÓN	1	1	0	0
DOSIS TOTAL (cat)	1	0	0	0
FRACCIONES	1	0	1	0
PTV VOLUMEN (cm3)	0	0	0	0
INMUNOTERAPIA	1	0	0	0

b) Clíniques i radiòmiques

	Correlational_corr	Mutual Informatio	Random Forest	ANOVA
ID	1	0	1	0
RESPUESTA POR TC A 1.5 MESES	1	0	0	0
EDAD	0	0	0	0
SEXO	1	0	0	0
TABAQUISMO	0	0	0	0
ENOLISMO	0	1	0	0
DM	0	0	0	0
HTA	1	0	0	0
DLP	0	1	0	1
HISTOLOGIA	0	0	0	0
ESTADIAJE CLÍNICO (T)	0	0	0	0
ESTADI CLÍNIC general	1	0	0	0
LOCALIZACIÓN	0	1	0	1
DOSIS TOTAL (cat)	1	0	0	0
FRACCIONES	1	0	1	0
PTV VOLUMEN (cm3)	0	0	0	0
INMUNOTERAPIA	1	1	1	0

GrayLevelNonUniformity	1	1	1	0
GrayLevelVariance	0	1	1	0
HighGrayLevelEmphasis	0	0	1	0
LargeDependenceEmphasis	0	1	0	0
geDependenceHighGrayLevelEmph	0	0	1	0
geDependenceLowGrayLevelEmph	0	0	0	0
LowGrayLevelEmphasis	1	1	0	0
SmallDependenceEmphasis	1	1	0	1
allDependenceHighGrayLevelEmph	0	0	1	0
allDependenceLowGrayLevelEmph	0	1	1	0
GrayLevelNonUniformity.1	1	1	1	1
rayLevelNonUniformityNormalize	0	1	0	1
GrayLevelVariance.1	0	1	0	0
HighGrayLevelRunEmphasis	0	0	1	0
LongRunEmphasis	0	1	0	0
LongRunHighGrayLevelEmphasis	0	0	1	0
LongRunLowGrayLevelEmphasis	1	0	1	0
LowGrayLevelRunEmphasis	1	0	0	0
RunEntropy	0	0	0	0
RunLengthNonUniformity	1	0	0	0
unLengthNonUniformityNormalize	1	1	0	1
RunPercentage	0	1	1	0
RunVariance	0	1	0	0
ShortRunEmphasis	1	1	0	0
ShortRunHighGrayLevelEmphasis	0	0	0	0
ShortRunLowGrayLevelEmphasis	1	0	1	0
GrayLevelNonUniformity.2	1	1	1	0
rayLevelNonUniformityNormalized	0	0	0	0
GrayLevelVariance.2	0	1	0	0
HighGrayLevelZoneEmphasis	0	0	1	0
LargeAreaEmphasis	1	1	0	0
LargeAreaHighGrayLevelEmphasis	0	1	0	0
LargeAreaLowGrayLevelEmphasis	0	0	1	0
LowGrayLevelZoneEmphasis	0	1	0	0
SizeZoneNonUniformity	1	0	1	0
izeZoneNonUniformityNormalized	0	1	1	1
SmallAreaEmphasis	0	1	0	1
SmallAreaHighGrayLevelEmphasis	0	0	1	0
SmallAreaLowGrayLevelEmphasis	0	1	0	1
ZoneEntropy	0	0	0	0
ZonePercentage	1	0	1	0
ZoneVariance	1	1	1	0
Busyness	1	0	1	0
Coarseness	1	1	0	1
Complexity	0	0	1	0
Contrast.1	1	1	1	0
Strength	1	1	0	0

Elongation	1	0	1	0
Flatness	1	0	1	0
LeastAxisLength	1	1	1	1
MajorAxisLength	1	1	1	0
Maximum2DDiameterColumn	1	1	1	0
Maximum2DDiameterRow	1	1	0	1
Maximum2DDiameterSlice	1	1	1	1
Maximum3DDiameter	1	1	1	1
MeshVolume	1	1	0	1
MinorAxisLength	1	1	1	0
Sphericity	0	1	1	0
SurfaceArea	1	1	0	0
SurfaceVolumeRatio	1	1	1	0
VoxelVolume	1	1	1	1
10Percentile	1	1	1	1
90Percentile	0	0	0	0
Energy	1	0	1	0
Entropy	0	0	0	0
InterquartileRange	0	1	0	1
Kurtosis	0	0	0	0
Maximum	0	1	0	1
MeanAbsoluteDeviation	0	1	0	1
Mean	0	0	1	0
Median	0	0	0	0
Minimum	0	1	1	1
Range	0	1	1	1
RobustMeanAbsoluteDeviation	0	1	0	0
RootMeanSquared	1	0	0	0
Skewness	0	0	0	0
TotalEnergy	1	0	1	0
Uniformity	0	0	1	0
Variance	0	1	0	0
Autocorrelation	0	0	1	0
ClusterProminence	0	1	0	0
ClusterShade	0	0	1	0
ClusterTendency	0	1	0	0
Contrast	1	0	0	0
Correlation	1	0	1	0
DifferenceAverage	1	1	0	0
DifferenceEntropy	1	0	1	0
DifferenceVariance	1	0	0	0
Id	1	0	0	0
Idm	1	0	1	0
Idmn	1	1	0	0
Idn	1	0	0	0
Imc1	1	1	0	1
Imc2	1	1	0	0
InverseVariance	1	0	0	0
JointAverage	0	0	0	0
JointEnergy	0	1	1	0
JointEntropy	0	0	0	0
MCC	0	0	0	0
MaximumProbability	0	0	0	0
SumAverage	0	0	0	0
SumEntropy	0	1	0	0
SumSquares	0	0	0	0
DependenceEntropy	1	0	0	0
DependenceNonUniformity	1	0	1	0
pendenceNonUniformityNormaliz	1	1	1	0
DependenceVariance	0	1	1	0
GrayLevelNonUniformity	1	1	1	0

c) Clíniques i metabolòmiques.

	Correlational_corr	Mutual Information	Random Forest	ANOVA
ID	1	0	1	0
RESPUESTA POR TC A 1.5 MESES	1	0	0	0
EDAD	0	0	0	0
SEXO	0	1	0	0
TABAQUISMO	0	0	0	0
ENOLISMO	1	0	0	0
DM	0	1	0	1
HTA	1	1	0	0
DLP	1	1	0	0
HISTOLOGIA	0	0	0	0
ESTADIAJE CLÍNICO (T)	1	0	0	0
ESTADI CLÍNIC general	1	0	0	0
LOCALIZACIÓN	1	1	0	0
DOSIS TOTAL (cat)	1	1	0	0
FRACCIONES	1	1	1	0
PTV VOLUMEN (cm3)	1	0	0	0
INMUNOTERAPIA	1	1	1	0
Pyruvic acid	0	1	1	1
Lactic acid	1	0	1	0
2-Hydroxyisobutyric acid	1	0	0	0
Glycolic acid	0	0	1	0
Alanine	0	1	0	0
2-HydroxyButyric acid	0	0	1	0
yl-2-oxobutyric acid (alphaketoisovaler	0	1	1	0
droxybutyric acid/3-hydroxyisobutyric	1	1	0	0
2-Hydroxyisovaleric acid	0	0	0	0
2-keto-3-methylvaleric acid	0	0	1	0
3-Hydroxyisovaleric acid	0	0	0	0
Valine	0	1	1	0
Benzoic acid	1	1	1	0
Ethanolamine	0	1	0	0
Leucine	0	1	1	0
Phosphoric acid	0	0	1	0
Glycerol	1	0	0	0
Ethylmalonic acid	0	1	0	0
Isoleucine	0	0	0	0
Proline	0	0	0	0
Glycine	0	1	1	0
Succinic acid	0	0	0	0
Glyceric acid	0	1	0	0
Fumaric acid	0	0	0	0
Serine	0	0	1	0
Threonine	0	1	0	1
Hydrocinnamic acid	0	0	1	0
Malic acid	0	0	0	0
d-Threitol	0	1	0	0
Methionine	0	0	1	0
Oxoproline	1	0	0	0
4-Hydroxyproline	1	0	0	0
Threonic acid	1	0	1	0
Erythronic acid	1	1	0	0
DL-2-Hydroxyglutaric acid	0	0	0	0
a-ketoglutaric acid	0	0	1	0
Glutamic acid	1	0	0	0
4-Hydroxybenzoic acid	0	0	1	0
Phenylalanine	0	1	1	0
Dodecanoic acid	1	1	0	0
d-Xylose	0	0	1	0
Taurine	0	1	0	0

d-Arabinose	1	1	0	0
d-Xylitol	1	0	0	0
d-Arabitol	0	0	0	0
Glycerol-1-phosphate	1	1	1	0
Glutamine	1	1	1	0
Xylonic acid	1	0	1	0
Ribonic acid	0	1	0	0
3-Phosphoglyceric acid	0	1	1	0
Ornithine	0	0	0	0
Citric acid	0	0	1	0
Tetradecanoic acid	1	1	0	0
Hippuric acid	0	0	0	0
Vanillylmandelic acid	0	1	0	0
4-hydroxyPhenyllactic acid	1	0	0	0
d-Fructose	0	0	1	0
d-Mannitol	1	0	1	0
d-Mannonic acid	1	1	0	0
d-Galactitol	0	0	1	0
Galacturonic acid	1	1	1	1
Galactonic acid	0	0	0	0
Saccharic acid	1	0	1	0
Indole-3-propanoic acid	0	0	1	0
Myo-Inositol	0	0	1	0
Uric acid	1	0	1	0
Sedoheptulose	1	0	1	0
Indolelactic acid	0	1	1	0
Linoleic acid	1	1	0	0
Oleic acid	0	0	1	0
Glucose 6-phosphate	0	0	0	0
d-Sucrose	1	0	0	0
Maltose	0	0	1	0
a-Tocopherol	0	0	0	0
Pyruvic acid .1	0	1	0	0
Lactic acid .1	0	0	1	0
2-Hydroxyisobutyric acid .1	1	1	0	0
Glycolic acid .1	0	0	0	0
Alanine .1	0	0	0	0
2-HydroxyButyric acid .1	0	1	0	0
l-2-oxobutyric acid (alphaketoisovaleri	0	0	0	0
roxybutyric acid/3-hydroxyisobutyric a	0	0	0	0
2-Hydroxyisovaleric acid .1	0	1	0	0
2-keto-3-methylvaleric acid .1	0	0	1	0
3-Hydroxyisovaleric acid .1	1	1	0	1
Valine .1	0	0	1	0
Benzoic acid .1	1	0	1	0
Ethanolamine .1	0	0	0	0
Leucine .1	0	1	1	0
Phosphoric acid .1	0	0	0	0
Glycerol .1	0	1	0	1
Ethylmalonic acid .1	0	0	0	0
Isoleucine .1	0	1	0	0
Proline .1	0	0	0	0
Glycine .1	0	0	1	0
Succinic acid .1	0	1	1	0
Glyceric acid .1	0	0	0	0
Fumaric acid .1	0	0	0	0
Serine .1	0	0	0	0
Threonine .1	0	0	0	0
Hydrocinnamic acid .1	0	1	1	0
Malic acid .1	0	1	0	0

d-Threitol .1	0	1	0	0
Methionine .1	0	1	1	0
Oxoproline .1	1	1	1	0
4-Hydroxyproline .1	1	0	0	0
Threonic acid .1	1	0	1	0
Erythronic acid .1	1	0	0	0
DL-2-Hydroxyglutaric acid .1	0	0	0	0
a-ketoglutaric acid .1	0	0	1	0
Glutamic acid .1	1	0	1	0
4-Hydroxybenzoic acid .1	0	0	0	0
Phenylalanine .1	0	0	0	0
Dodecanoic acid .1	0	0	0	0
d-Xylose .1	0	0	0	0
Taurine .1	0	0	1	0
d-Arabinose .1	1	1	0	0
d-Xylitol .1	0	1	1	0
d-Arabitol .1	0	1	0	0
Glycerol-1-phosphate .1	0	0	1	0
Glutamine .1	1	1	0	1
Xylonic acid .1	0	0	1	0
Ribonic acid .1	0	1	0	0
3-Phosphoglyceric acid .1	0	0	0	0
Ornithine .1	0	1	0	0
Citric acid .1	0	0	1	0
Tetradecanoic acid .1	0	1	0	0
Hippuric acid .1	0	1	0	0
Vanillylmandelic acid .1	0	1	0	0
4-hydroxyPhenyllactic acid .1	0	0	0	0
d-Fructose .1	0	0	1	0
d-Mannitol .1	0	0	0	0
d-Mannonic acid .1	0	1	0	1
d-Galactitol .1	0	1	0	1
Galacturonic acid .1	0	0	0	0
Galactonic acid .1	0	0	1	0
Saccharic acid .1	0	1	1	0
Indole-3-propanoic acid .1	0	0	0	0
Myo-Inositol .1	0	1	1	0
Uric acid .1	1	1	0	0
Sedoheptulose .1	0	1	0	0
Indolelactic acid .1	0	0	1	0
Linoleic acid .1	0	0	1	0
Oleic acid .1	0	0	0	0
Glucose 6-phosphate .1	0	0	0	0
d-Sucrose .1	0	0	0	0
Maltose.1	0	1	1	0
a-Tocopherol .1	0	1	0	0

d) Clíniques, radiòmiques i metabolòmiques

	Correlational_corr	Mutual Information	Random Forest	ANOVA
ID	1	0	0	0
ESPUESTA POR TC A 1.5 MESI	1	0	1	0
EDAD	0	0	0	0
SEXO	0	1	0	0
TABAQUISMO	0	0	0	0
ENOLISMO	1	0	0	0
DM	0	0	1	0
HTA	1	1	0	0
DLP	1	0	0	0
HISTOLOGIA	0	1	0	1
ESTADIAJE CLÍNICO (T)	1	0	0	0
ESTADI CLÍNIC general	1	0	0	0
LOCALIZACIÓN	0	1	0	0
DOSIS TOTAL (cat)	1	0	0	0
FRACCIONES	1	1	1	0
PTV VOLUMEN (cm3)	1	0	0	0
INMUNOTERAPIA	1	1	0	1

Pyruvic acid	0	1	0	1
Lactic acid	1	0	1	0
2-Hydroxyisobutyric acid	1	0	0	0
Glycolic acid	0	0	0	0
Alanine	0	1	1	0
2-HydroxyButyric acid	0	1	0	0
xobutyric acid (alphaketoisc	0	1	1	0
butyric acid/3-hydroxyisobu	1	1	0	0
2-Hydroxyisovaleric acid	0	0	0	0
2-keto-3-methylvaleric acid	0	0	0	0
3-Hydroxyisovaleric acid	0	1	0	0
Valine	0	1	0	0
Benzoic acid	1	1	1	1
Ethanolamine	0	1	0	0
Leucine	0	1	1	0
Phosphoric acid	0	0	1	0
Glycerol	1	0	0	0
Ethylmalonic acid	0	1	1	0
Isoleucine	0	0	1	0
Proline	0	0	0	0
Glycine	0	1	1	1
Succinic acid	0	0	0	0
Glyceric acid	0	1	1	0
Fumaric acid	0	0	0	0
Serine	0	0	0	0
Threonine	0	1	0	0
Hydrocinnamic acid	0	0	1	0
Malic acid	0	0	1	0
d-Threitol	0	1	0	0
Methionine	0	0	0	0
Oxoproline	1	0	0	0
4-Hydroxyproline	1	0	1	0
Threonic acid	1	1	1	0
Erythronic acid	1	1	0	0
DL-2-Hydroxyglutaric acid	0	0	0	0
a-ketoglutaric acid	0	0	1	0
Glutamic acid	1	0	0	0
4-Hydroxybenzoic acid	0	0	0	0
Phenylalanine	0	0	1	0
Dodecanoic acid	1	1	1	0
d-Xylose	0	0	0	0
Taurine	0	1	0	0
d-Arabinose	1	1	0	0
d-Xylitol	1	0	0	0
d-Arabitol	0	0	0	0
Glycerol-1-phosphate	1	0	0	0
Glutamine	1	1	0	0
Xylonic acid	1	0	1	0
Ribonic acid	0	1	0	0
3-Phosphoglyceric acid	0	1	1	0
Ornithine	0	0	0	0
Citric acid	0	0	1	0

Tetradecanoic acid	1	1	0	0
Hippuric acid	0	0	0	0
Vanillylmandelic acid	0	1	1	0
4-hydroxyPhenyllactic acid	1	1	0	0
d-Fructose	0	0	0	0
d-Mannitol	1	0	0	0
d-Mannonic acid	1	0	0	0
d-Galactitol	0	0	1	0
Galacturonic acid	1	1	0	0
Galactonic acid	0	0	0	0
Saccharic acid	0	0	1	0
Indole-3-propanoic acid	0	0	1	0
Myo-Inositol	0	0	1	0
Uric acid	1	1	0	0
Sedoheptulose	1	1	1	0
Indolelactic acid	0	0	1	0
Linoleic acid	1	1	1	0
Oleic acid	0	1	0	0
Glucose 6-phosphate	0	0	1	0
d-Sucrose	1	0	1	0
Maltose	0	0	1	0
a-Tocopherol	0	1	0	0
Pyruvic acid .1	0	1	0	0
Lactic acid .1	0	0	1	0
2-Hydroxyisobutyric acid .1	1	1	0	0
Glycolic acid .1	0	0	1	0
Alanine .1	0	0	0	0
2-HydroxyButyric acid .1	0	1	0	1
xobutyric acid (alphaketoisov	0	1	0	0
butyric acid/3-hydroxyisobut	0	1	0	0
2-Hydroxyisovaleric acid .1	0	1	1	0
-keto-3-methylvaleric acid .	0	0	0	0
3-Hydroxyisovaleric acid .1	1	1	0	0
Valine .1	0	0	1	0
Benzoic acid .1	1	0	0	0
Ethanolamine .1	0	0	1	0
Leucine .1	0	1	1	0
Phosphoric acid .1	0	0	1	0
Glycerol .1	0	0	0	0
Ethylmalonic acid .1	0	0	0	0
Isoleucine .1	0	1	0	0
Proline .1	0	0	0	0
Glycine .1	0	1	0	0
Succinic acid .1	0	1	1	0
Glyceric acid .1	1	0	1	0
Fumaric acid .1	0	0	1	0
Serine .1	0	0	0	0
Threonine .1	0	0	0	0
Hydrocinnamic acid .1	0	1	1	0
Malic acid .1	0	1	0	0
d-Threitol .1	1	1	1	0
Methionine .1	0	1	0	0
Oxoproline .1	1	1	1	0
4-Hydroxyproline .1	1	0	0	0
Threonic acid .1	1	1	0	0

Erythronic acid .1	1	0	1	0
DL-2-Hydroxyglutaric acid .1	0	0	1	0
a-ketoglutaric acid .1	0	0	1	0
Glutamic acid .1	1	0	0	0
4-Hydroxybenzoic acid .1	0	0	1	0
Phenylalanine .1	0	0	0	0
Dodecanoic acid .1	0	0	0	0
d-Xylose .1	0	0	1	0
Taurine .1	0	0	0	0
d-Arabinose .1	1	1	0	0
d-Xylitol .1	0	1	0	0
d-Arabitol .1	0	1	1	1
Glycerol-1-phosphate .1	1	0	0	0
Glutamine .1	1	1	1	0
Xylonic acid .1	0	0	1	0
Ribonic acid .1	0	1	1	0
3-Phosphoglyceric acid .1	0	1	0	0
Ornithine .1	0	1	1	1
Citric acid .1	0	0	0	0
Tetradecanoic acid .1	0	0	0	0
Hippuric acid .1	0	0	0	0
Vanillylmandelic acid .1	0	1	0	0
4-hydroxyPhenyllactic acid .1	0	0	1	0
d-Fructose .1	0	0	0	0
d-Mannitol .1	0	0	1	0
d-Mannonic acid .1	0	1	1	0
d-Galactitol .1	0	1	0	0
Galacturonic acid .1	0	0	0	0
Galactonic acid .1	0	0	1	0
Saccharic acid .1	0	1	1	0
Indole-3-propanoic acid .1	0	0	1	0
Myo-Inositol .1	0	1	1	0
Uric acid .1	1	1	1	0
Sedoheptulose .1	0	1	0	0
Indolelactic acid .1	0	0	0	0
Linoleic acid .1	0	0	0	0
Oleic acid .1	0	0	0	0
Glucose 6-phosphate .1	0	0	0	0
d-Sucrose .1	0	0	1	0
Maltose.1	0	1	1	0
a-Tocopherol .1	0	1	1	0
Elongation	1	1	1	1
Flatness	1	1	1	0
LeastAxisLength	1	1	1	0
MajorAxisLength	1	1	1	0
Maximum2DDiameterColumn	1	1	1	1
Maximum2DDiameterRow	1	1	0	0
Maximum2DDiameterSlice	1	1	1	0
Maximum3DDiameter	1	1	0	0
MeshVolume	1	1	0	0
MinorAxisLength	1	1	0	0
Sphericity	1	1	1	0
SurfaceArea	1	1	0	0
SurfaceVolumeRatio	1	0	1	0
VoxelVolume	1	1	0	0

10Percentile	0	1	1	0
90Percentile	0	0	0	0
Energy	1	0	0	0
Entropy	0	1	0	1
InterquartileRange	0	1	0	0
Kurtosis	0	0	1	0
Maximum	0	1	0	0
MeanAbsoluteDeviation	0	1	1	0
Mean	0	0	1	0
Median	0	0	1	0
Minimum	0	1	1	0
Range	0	0	0	0
obustMeanAbsoluteDeviatio	0	1	0	0
RootMeanSquared	0	0	0	0
Skewness	0	0	0	0
TotalEnergy	1	0	0	0
Uniformity	0	1	0	0
Variance	0	1	0	0
Autocorrelation	0	0	0	0
ClusterProminence	1	1	0	0
ClusterShade	1	0	0	0
ClusterTendency	0	1	0	0
Contrast	1	1	0	0
Correlation	1	0	0	0
DifferenceAverage	1	1	0	0
DifferenceEntropy	0	0	0	0
DifferenceVariance	1	0	1	0
Id	0	0	0	0
Idm	0	0	0	0
Idmn	1	1	0	0
Idn	1	0	0	0
Imc1	1	1	0	0
Imc2	0	0	0	0
InverseVariance	1	0	0	0
JointAverage	0	0	1	0
JointEnergy	0	0	0	0
JointEntropy	0	0	0	0
MCC	0	0	0	0
MaximumProbability	0	0	0	0
SumAverage	0	0	1	0
SumEntropy	0	0	0	0
SumSquares	0	1	1	0
DependenceEntropy	1	0	0	0
DependenceNonUniformity	1	1	0	0
ndenceNonUniformityNorm	1	0	1	0
DependenceVariance	0	1	1	0
GrayLevelNonUniformity	1	1	1	0
GrayLevelVariance	0	1	0	0
HighGrayLevelEmphasis	0	0	0	0
LargeDependenceEmphasis	0	1	1	0
ependenceHighGrayLevelEm	0	0	1	0
ependenceLowGrayLevelEm	1	0	1	0
LowGrayLevelEmphasis	0	1	0	0

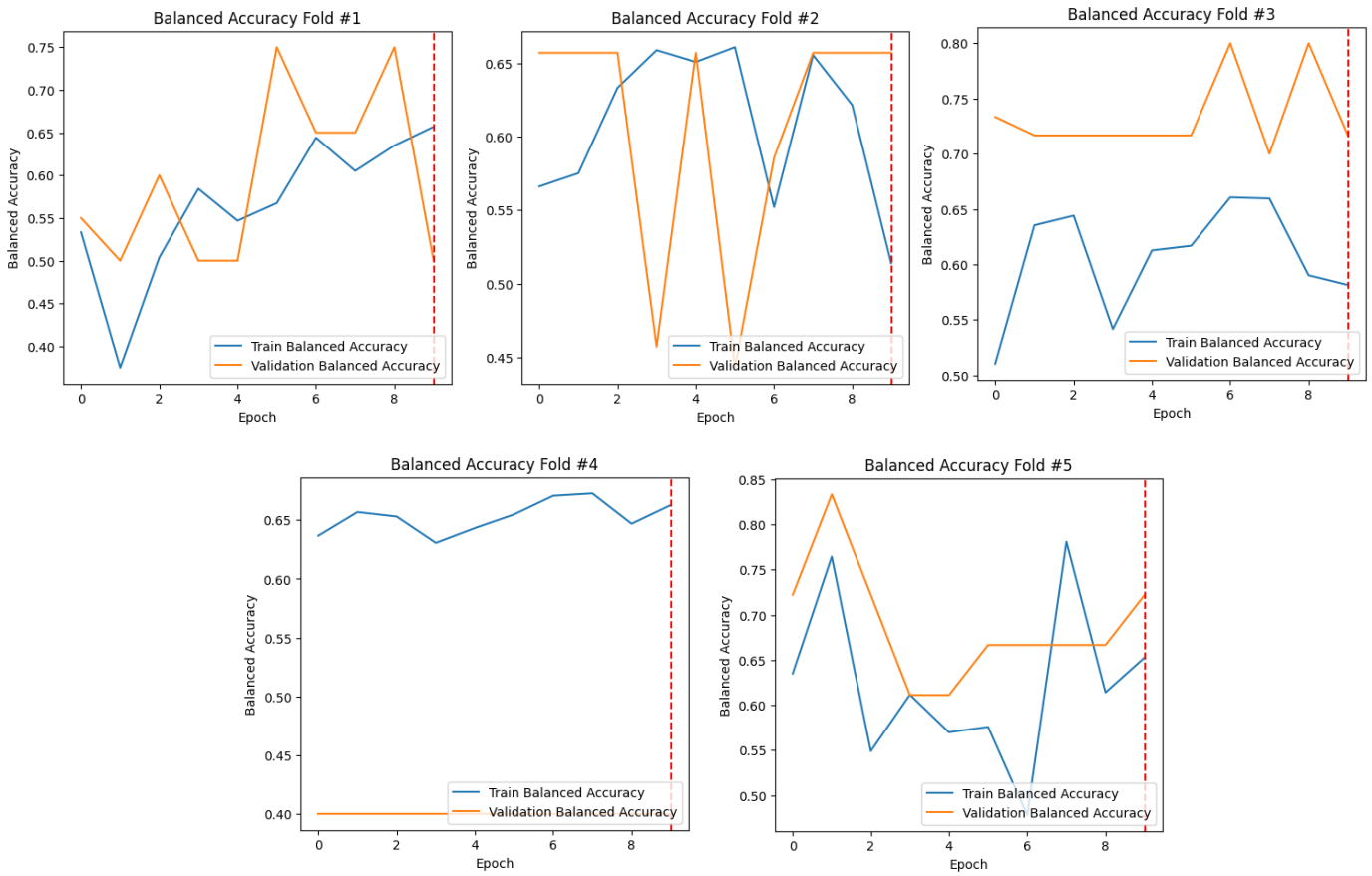
SmallDependenceEmphasis	1	0	1	0
ependenceHighGrayLevelEn	0	0	1	0
ependenceLowGrayLevelEm	0	1	0	0
GrayLevelNonUniformity.1	1	1	0	0
yLevelNonUniformityNormal	0	1	1	0
GrayLevelVariance.1	0	1	0	0
HighGrayLevelRunEmphasis	0	0	1	0
LongRunEmphasis	0	1	1	1
ngRunHighGrayLevelEmphas	0	0	0	0
ongRunLowGrayLevelEmphas	0	0	0	0
LowGrayLevelRunEmphasis	0	1	0	0
RunEntropy	0	0	0	0
RunLengthNonUniformity	1	0	0	0
engthNonUniformityNorma	0	1	1	0
RunPercentage	0	1	1	0
RunVariance	0	1	0	0
ShortRunEmphasis	0	1	1	1
ortRunHighGrayLevelEmpha	0	0	0	0
ortRunLowGrayLevelEmpha	0	1	0	0
GrayLevelNonUniformity.2	1	1	0	0
LevelNonUniformityNormali	0	0	0	0
GrayLevelVariance.2	0	0	0	0
HighGrayLevelZoneEmphasis	0	0	1	0
LargeAreaEmphasis	1	1	1	1
geAreaHighGrayLevelEmpha	1	1	0	0
geAreaLowGrayLevelEmpha	1	0	0	0
LowGrayLevelZoneEmphasis	0	0	1	0
SizeZoneNonUniformity	1	1	1	0
ZoneNonUniformityNormali	0	1	1	0
SmallAreaEmphasis	0	1	1	0
allAreaHighGrayLevelEmpha	0	0	1	0
allAreaLowGrayLevelEmpha	0	0	0	0
ZoneEntropy	0	0	0	0
ZonePercentage	1	0	1	0
ZoneVariance	1	1	1	0
Busyness	1	0	1	0
Coarseness	1	1	0	0
Complexity	0	0	0	0
Contrast.1	1	1	1	0
Strength	1	1	0	0

Annex II

Es presenta l'evolució del model de CNN optimitzat amb el càlcul del valor de *balanced accuracy*.

a) Clíniques

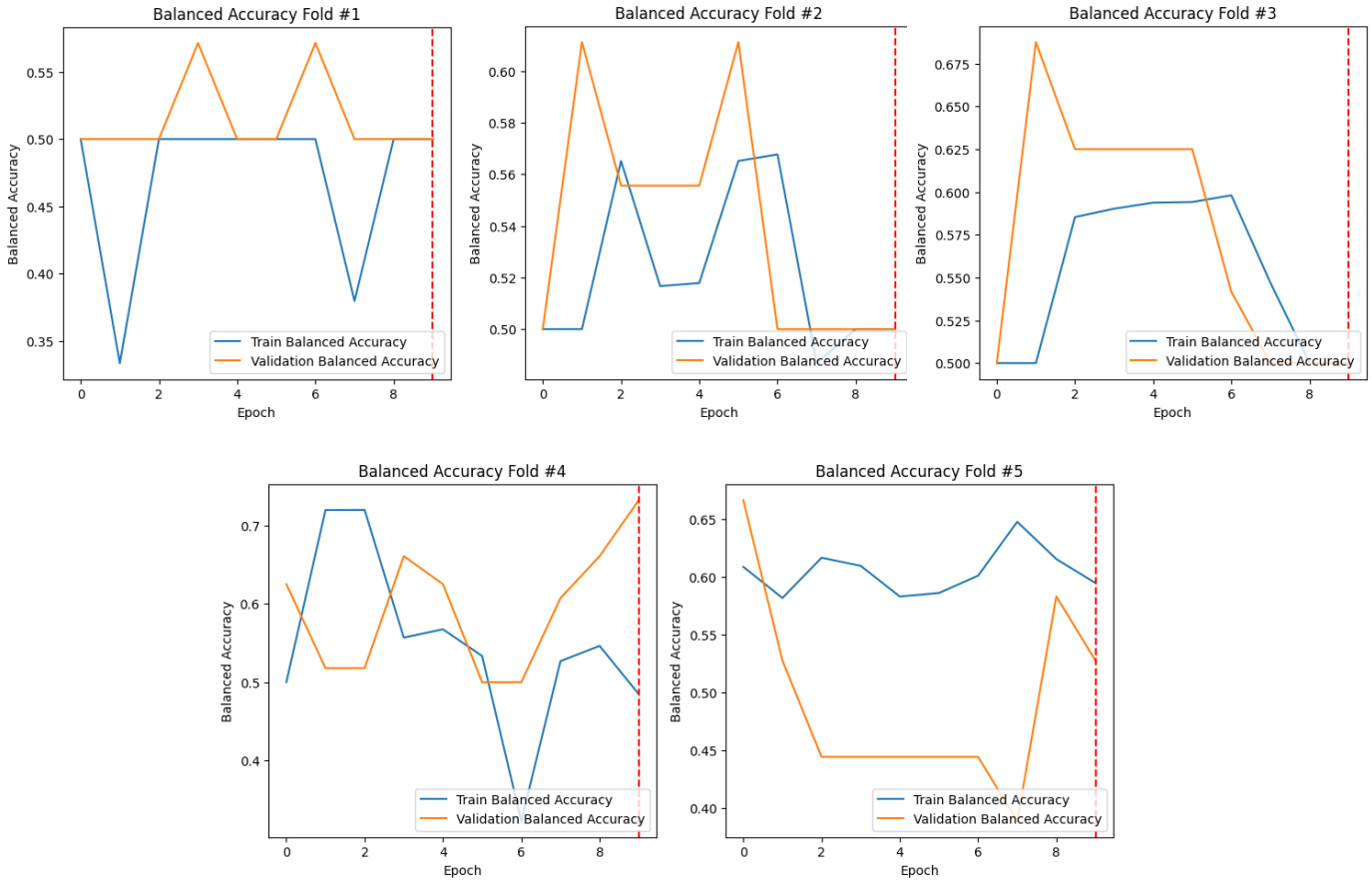
La primera iteració, mostra una tendència decreixent a mesura que avancen les èpoques, amb pocs signes de convergència amb la precisió d'entrenament. En la segona, hi ha una fluctuació notable en la precisió de validació, amb un pic al començament seguit d'una caiguda i una posterior recuperació. Aquest patró volàtil podria ser un indicador de inestabilitat en el procés d'aprenentatge, potencialment causat per una taxa d'aprenentatge massa alta o per la falta de regularització adequada. En el tercer, s'observa un patró decreixent tant en la precisió d'entrenament com de validació, amb una divergència notable cap a les últimes èpoques. Això pot ser un senyal que el model necessita més èpoques per estabilitzar-se o que les característiques de les dades en aquest plec són particularment difícils de modelar. En el quart es manté una estabilitat, més aviat millora del model. En canvi en el cinquè, la precisió de validació cau dràsticament després de les primeres èpoques i després mostra una recuperació lleugera, però no arriba al nivell inicial. Aquest comportament suggeriria revisar la taxa d'aprenentatge o aplicar mètodes de parada primerenca per prevenir una caiguda tan dràstica.



Grup de figures 1. Evolució del model CNN quan s'utilitzen dades clíniques.

c) Clíniques i radiòmiques.

Tant la primera com la quarta iteració, es veu un augment inicial sostingut de la precisió balancejada tant per a les dades d'entrenament com de validació, amb un tancament alt just abans de l'aturada anticipada (indicat per la línia vertical roja puntejada). En la segona, s'observa una caiguda abrupta en la precisió de validació després d'un breu augment inicial. En la tercera iteració, la precisió de validació mostra volatilitat notable, amb fluctuacions significatives que, no obstant això, mantenen un nivell relativament alt comparat amb la precisió d'entrenament. Això pot indicar que el model té una capacitat raonable de generalització, tot i la variabilitat observada. En la cinquena, hi ha una estabilitat inicial en la precisió de validació seguida d'un descens i una millora notable just abans de l'aturada anticipada. Aquest comportament pot suggereix que ajustaments tardans al model o al procés d'entrenament podrien millorar la seva capacitat de generalització.

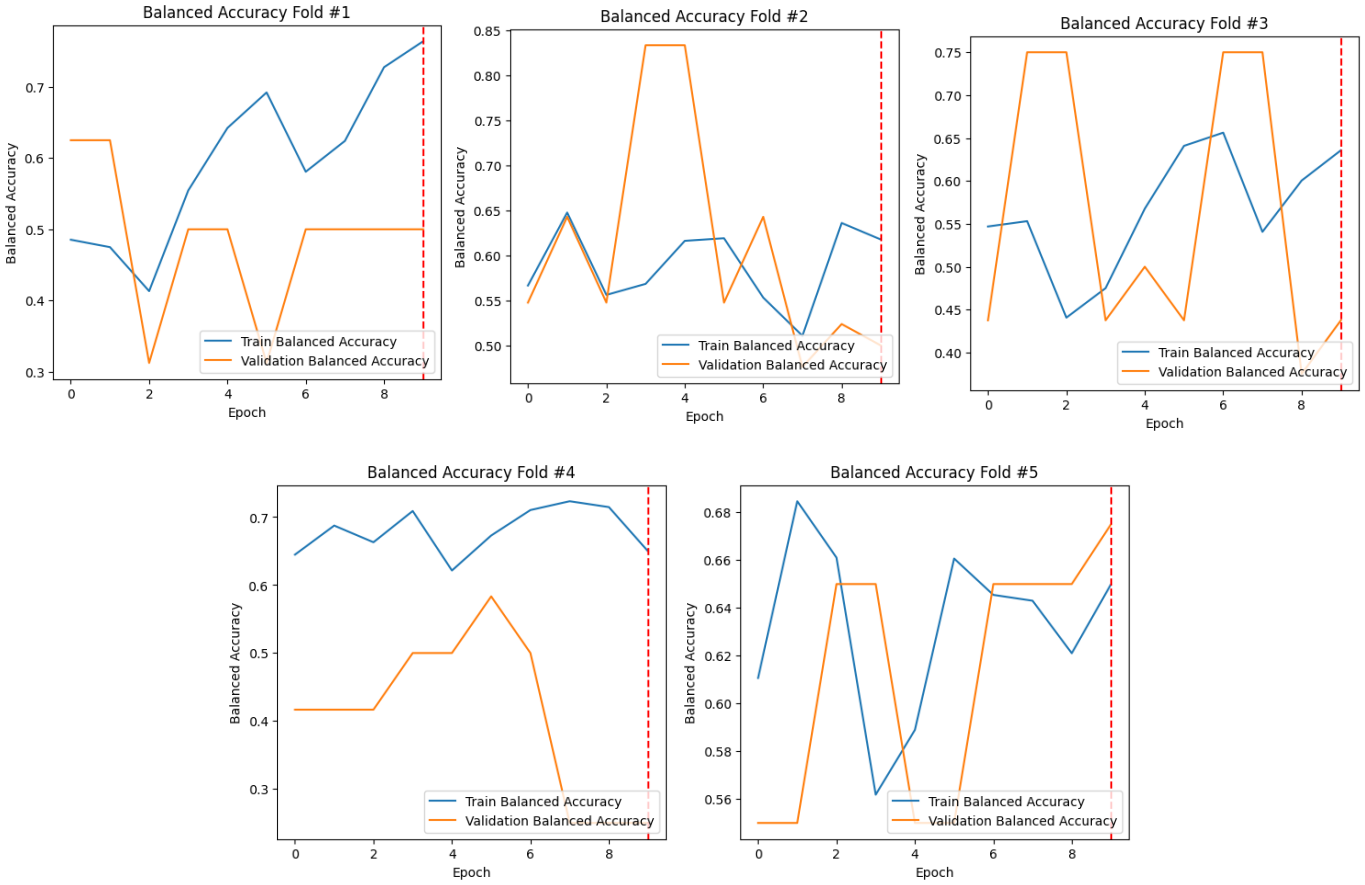


Grup de figures 2. Evolució del model CNN quan s'utilitzen dades clíniques i radiòmiques.

d) Clíniques i metabolòmiques.

En les iteracions 1 i 4, es veu un augment inicial sostingut de la precisió balancejada tant per a les dades d'entrenament com de validació, amb un tancament alt just abans de l'aturada anticipada. Aquest patró suggereix un bon ajustament del model, amb poca evidència de sobreajustament. En canvi, en la segona s'observa, una caiguda abrupta en la precisió de validació després d'un breu augment inicial. L'aturada anticipada es produeix molt primerenca, possiblement a causa d'una mala generalització o un ajust inicial no òptim del model. Pot ser indicatiu de la necessitat de revisar els paràmetres del model o la necessitat de més dades o una millor pre-processament de les dades. En la

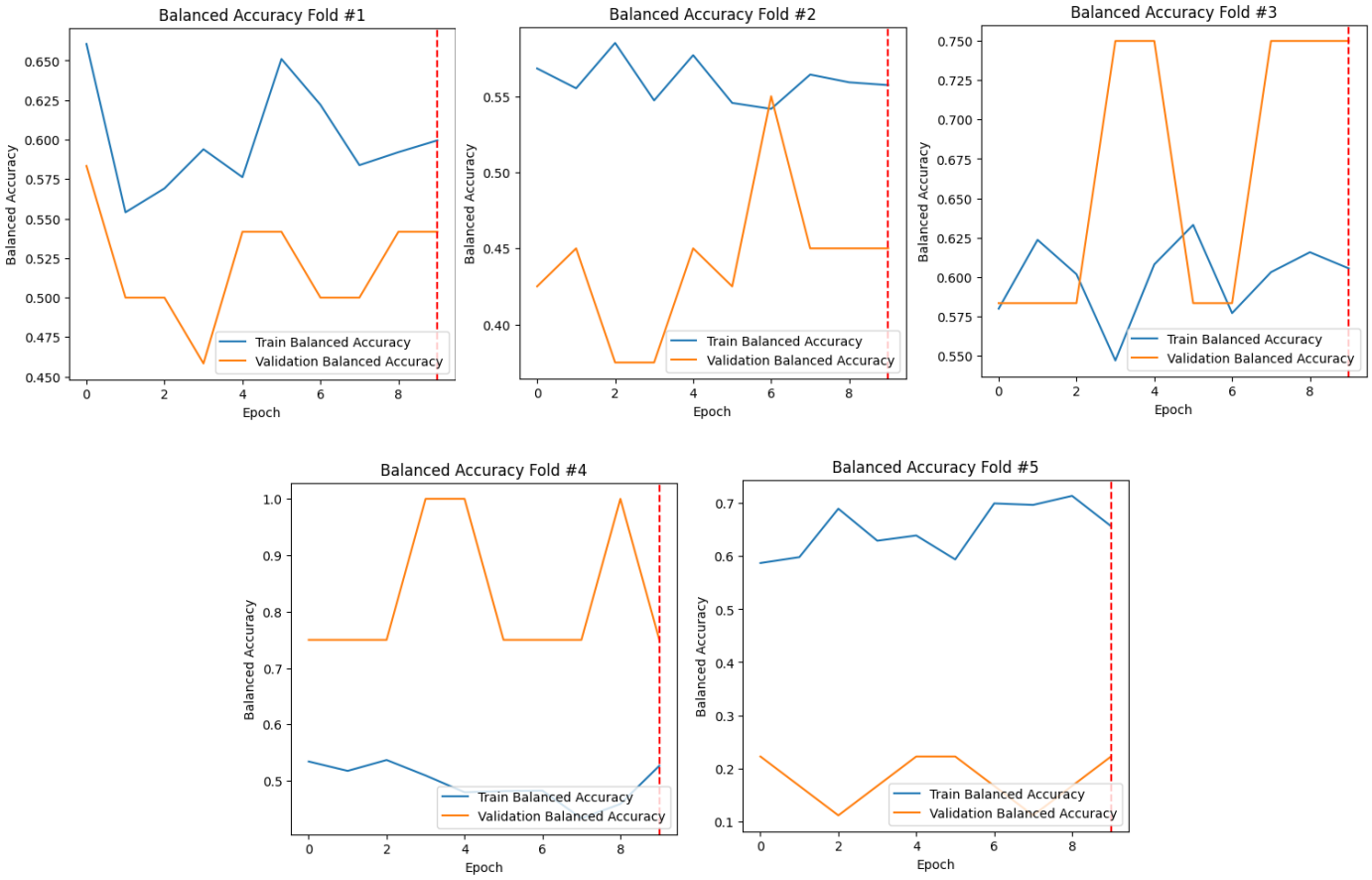
tercera, la precisió de validació mostra volatilitat notable, amb fluctuacions significatives que, no obstant això, mantenen un nivell relativament alt comparat amb la precisió d'entrenament. Això pot indicar que el model té una capacitat raonable de generalització, tot i la variabilitat observada. Finalment en la 5a iteració, es veu una estabilitat inicial en la precisió de validació seguida d'un descens i una millora notable just abans de l'aturada anticipada.



Grup de figures 3. Evolució del model CNN quan s'utilitzen dades clíniques i metabolòmiques.

e) Clíniques, metabolòmiques i radiòmiques.

En les iteracions 1 i 4, hi ha un augment sostingut en la precisió d'entrenament i de validació en les primeres iteracions. En ambdós casos, la precisió de validació mostra un pic alt just abans de l'aturada anticipada (*early stopping*), indicant un bon ajustament del model sense sobreajustament aparent. En les 2 i 3, es mostra una caiguda abrupta en la precisió de validació després d'un breu augment. Malgrat les oscil·lacions, la precisió de validació manté un nivell relativament alt comparat amb l'entrenament, suggerint una capacitat raonable del model per generalitzar malgrat la volatilitat. En la iteració 5, inicialment estable, la precisió de validació pateix un descens notable, després millora just abans de *l'early stopping*. Aquest comportament podria indicar que ajustaments tardans en el procés de formació podrien haver millorat la capacitat de generalització del model.



Grup de figures 4. Evolució del model CNN quan s'utilitzen dades clíniques, metabolòmiques i radiòmiques.

Annex III

Es presenta el codi utilitzat per el tractament de les dades, desenvolupament i validació dels models.

a) Imports

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import mutual_info_regression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.datasets import make_classification
from sklearn.feature_selection import f_classif, SelectKBest
from sklearn.feature_selection import SelectFromModel
import matplotlib.pyplot as plt
from keras.models import Sequential
from keras.layers import Conv1D, MaxPooling1D, Flatten, Dense
from keras.preprocessing.image import ImageDataGenerator
from sklearn.metrics import confusion_matrix, recall_score,
balanced_accuracy_score
from keras.optimizers import Adam
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix,
ConfusionMatrixDisplay
from imblearn.over_sampling import SMOTE
from keras import backend as K
from keras.models import Sequential, Model
from keras.layers import Conv1D, MaxPooling1D, Dense, Flatten
from sklearn.model_selection import KFold
```

b) Pujar les dades i processat de dades

```
df = pd.read_excel('cliniques_metabolomiques_radiòmiques.xlsx')
```

```
le = LabelEncoder()
df['ESTADI CLÍNIC general'] = le.fit_transform(df['ESTADI CLÍNIC
general'])
df.head()
```

```
df = df.dropna()

# Crear instancia StandardScaler
scaler = StandardScaler()

# Escalar totes les columnas menys la resposta
```

```
column1 = df['RESPUESTA POR TC A 1.5 MESES']
df_scaled = df.drop('RESPUESTA POR TC A 1.5 MESES', axis=1)
df_scaled = pd.DataFrame(scaler.fit_transform(df_scaled),
columns=df_scaled.columns)
df_scaled.insert(list(df.columns).index('RESPUESTA POR TC A 1.5 MESES'),
'RESPUESTA POR TC A 1.5 MESES', column1)
```

c) Selecció de característiques

```
X = df.drop('RESPUESTA POR TC A 1.5 MESES', axis=1)
y = df['RESPUESTA POR TC A 1.5 MESES']
y = y.astype(int)
```

```
df = df.apply(pd.to_numeric, errors='coerce')

corr_with_target = df.corrwith(df['RESPUESTA POR TC A 1.5 MESES']).abs()

high_corr_features = corr_with_target[corr_with_target >
0.15].index.tolist()
print(high_corr_features)

df_corr = df[high_corr_features]
col_reduced = df_corr.columns

print(col_reduced.to_list())
```

```
   mi = mutual_info_regression(X, y)

mi_scores = pd.DataFrame(mi, index=X.columns, columns=['MI
Scores']).sort_values(by='MI Scores', ascending=False)

selected_features = mi_scores[mi_scores['MI Scores'] > 0.02]
no_selected_features = mi_scores[mi_scores['MI Scores'] < 0.02]
selected_features = selected_features.transpose()

column_names = df.columns.tolist()

# Create a Random Forest classifier
rf = RandomForestClassifier(n_estimators=100)

# Train the classifier on the data
rf.fit(X, y)

# Select features based on the importance given by the Random Forest
selector = SelectFromModel(rf, prefit=True)
selected_feature_indices = selector.get_support(indices=True)
```

```
selected_feature_names = [column_names[i] for i in
selected_feature_indices]

# Transform the dataset to include only the important features
X_important = selector.transform(X)

# Create the DF with the selected features
X_important_df = pd.DataFrame(X_important,
columns=selected_feature_names)

print("Reduced dataset shape:", X_important_df.columns.to_list())
```

```
# Compute ANOVA F-value for the features
f_values, p_values = f_classif(X, y)
# Select a certain number of features based on the smallest p-values
# For example, to select the top 10 features
selector = SelectKBest(f_classif, k='all').fit(X, y)
# Transform X to the selected features
X_selected = selector.transform(X)

# If you want to see which features were selected:
selected_features_bool = selector.get_support()
selected_features_anova = [column for (column, selected) in
zip(df.columns, selected_features_bool) if selected]

#print(selected_features_anova)
X_important_Anova_df = pd.DataFrame(X_selected,
columns=selected_features_anova)

filtered_features_p_values = [(feature, p_value) for feature, p_value in
zip(selected_features, p_values[selected_features_bool]) if p_value <
0.1]
df_filtered_features = pd.DataFrame(filtered_features_p_values,
columns=['Feature', 'P-Value'])

# To print the selected features and their corresponding p-values
print("Selected features and their p-values:")
for feature, p_value in zip(selected_features,
p values[selected features bool]):
 if p_value < 0.1:
   print(f"{feature}: p-value = {p_value}")

features_presence = {
 "Correlational_corr":
df.columns.isin(df_corr.columns).astype(int),
 "Mutual Information":
df.columns.isin(selected_features.columns).astype(int),
 "Random Forest": df.columns.isin(X_important_df.columns).astype(int),
 "ANOVA": df.columns.isin(df_filtered_features['Feature']).astype(int)
```

```
}
#print(features_presence)
presence_absence_table = pd.DataFrame(features_presence,
index=df.columns)

print("Taula de presència-absència:")
presence_absence_table.head()
presence_absence_table.to_excel('presence_absence.xlsx')
```

```
def seleccionar_variables(df):
 # Select features with ones in three or more
 return df[df.sum(axis=1) >= 3].index.tolist()

# Use the function
variables_seleccionades = seleccionar_variables(presence_absence_table)

print(variables_seleccionades)
   print(len(variables_seleccionades))
```

d) Creació dels models

```
file_path_data_adapted = 'variables_amb_target.xlsx'

# Load the Excel file
dades_model = pd.read_excel(file_path_data_adapted, engine='openpyxl')

columnes_a_conservar = variables_seleccionades
columnes_a_conservar.append('RESPUESTA POR TC A 1.5 MESES')

# Select the columns to keep.

dades_model = df[variables_seleccionades]
dades_model.head()
```

```
y = dades_model['RESPUESTA POR TC A 1.5 MESES']
x = dades_model.drop('RESPUESTA POR TC A 1.5 MESES', axis=1)

X_model = dades_model.drop('RESPUESTA POR TC A 1.5 MESES', axis=1)
print(X_model)
y_model = dades_model['RESPUESTA POR TC A 1.5 MESES']
print(y model)
y_model_convertida = [0 if x == 1 or x == 2 else 1 for x in y_model]
X_train, X_val, y_train, y_val = train_test_split(X_model,
y_model_convertida, test_size=0.2, random_state=42)
```

e) Correlació

```
# Calculate the correlation matrix
correlation_matrix = dades_model.corr()

# Select the row corresponding to 'RESPUESTA POR TC A 1.5 MESES' (or your
column name)
correlation_with_response = correlation_matrix.loc['RESPUESTA POR TC A
1.5 MESES']

# Remove the 'RESPUESTA POR TC A 1.5 MESES' column from the correlation
series
correlation_with_response = correlation_with_response.drop('RESPUESTA POR
TC A 1.5 MESES')

# Create a horizontal bar chart of the correlation of each variable with
'RESPUESTA POR TC A 1.5 MESES'
plt.figure(figsize=(10, 6))
sns.barplot(x=correlation_with_response.values,
y=correlation_with_response.index, palette='coolwarm')

# Adjust the title and axis names
plt.title('Correlation of Variables with "RESPUESTA POR TC A 1.5 MESES"')
plt.xlabel('Correlation')
plt.ylabel('Variable')

# Set the x-axis limits to only show from -0.5 to 0.5
plt.xlim(-0.5, 0.5)

# Display the chart
plt.show()
```

f) Taula de presència

```
features_presence = {
    "Correlational_corr": df.columns.isin(df_corr.columns).astype(int),
    "Mutual Information":
df.columns.isin(selected_features.columns).astype(int),
    "Random Forest": df.columns.isin(X_important_df.columns).astype(int),
    "ANOVA": df.columns.isin(df_filtered_features['Feature']).astype(int)
}
#print(features_presence)
presence_absence_table = pd.DataFrame(features_presence,
index=df.columns)

print("Taula de presència-absència:")
presence_absence_table.head()
```

```
def seleccionar_variables(df):
    # Select rows that have a sum of 3 or more across specified columns
    return df[df.sum(axis=1) >= 3].index.tolist()

# Use the function to get the list of selected variables
variables_seleccionadas = seleccionar_variables(presence_absence_table)

print(variables_seleccionadas)
print(len(variables_seleccionadas))
```

g) Random Forest

```
file_path_data_adapted = 'variables_amb_target.xlsx'

# Load the Excel file
dades_model = pd.read_excel(file_path_data_adapted, engine='openpyxl')
```

```
# Initialize SMOTE with a random state for reproducibility
smote = SMOTE(random_state=20)
X train smote, y train smote = smote.fit resample(X train, y train)

# Create the RandomForestClassifier model
model = RandomForestClassifier(n_estimators=100, random_state=20)

# Train the model with the balanced data
model.fit(X_train_smote, y_train_smote)

# Make predictions using the validation data
y_pred = model.predict(X_val)

# Evaluate the model using accuracy
accuracy = accuracy_score(y_val, y_pred)
print(f'Model accuracy: {accuracy:.2f}')

# Generate the confusion matrix
conf_matrix = confusion_matrix(y_val, y_pred)

# Plot the confusion matrix using heatmap
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap='Blues',
xticklabels=model.classes_, yticklabels=model.classes_,
annot_kws={"size": 16})
plt.title('Confusion Matrix')
plt.ylabel('Actual')
plt.xlabel('Prediction')
plt.show()
```

h) CNN

```
def balanced_accuracy(y_true, y_pred):

    # Threshold predictions to get the predicted binary class (0 or 1)
    threshold = 0.5
    y_pred_thresholded = K.cast(K.greater(y_pred, threshold), K.floatx())

    # Calculate True Positives, False Positives, True Negatives, and
False Negatives
    tp = K.sum(y_true * y_pred_thresholded)
    tn = K.sum((1 - y_true) * (1 - y_pred_thresholded))
    fp = K.sum((1 - y_true) * y_pred_thresholded)
    fn = K.sum(y_true * (1 - y_pred_thresholded))

    # Calculate Sensitivity (Recall) and Specificity
    sensitivity = tp / (tp + fn + K.epsilon())
    specificity = tn / (tn + fp + K.epsilon())

    # Balanced Accuracy is the average of Sensitivity and Specificity
    balanced_acc = (sensitivity + specificity) / 2
    return balanced_acc
```

```
model = Sequential()
model.add(Conv1D(filters=70, kernel_size=3, activation='relu',
input_shape=(30, 1)))  # Updated input_shape
model.add(MaxPooling1D(pool_size=2))
model.add(Flatten())
model.add(Dense(20, activation='relu'))
model.add(Dense(1))

model.compile(optimizer='adam', loss='binary_crossentropy', metrics =
balanced_accuracy)
```

```
history_list = []
stop_epochs = []
conf_matrices = []

# Convertir X_model a DataFrame si no lo es
if isinstance(X model, list):
    X_model = pd.DataFrame(X_model)

# Convertir y_model_convertida a Series si no lo es
if isinstance(y model convertida, list):
    y_model_convertida = pd.Series(y_model_convertida)
```

```
kf = KFold(n_splits=5, shuffle=True, random_state=42)
all_classes = np.unique(y_model_convertida)

for train_index, test_index in kf.split(X_model):
    X_train, X_test = X_model.iloc[train_index], X_model.iloc[test_index]
    y_train, y_test = y_model_convertida.iloc[train_index],
y_model_convertida.iloc[test_index]

    # Apply SMOTE to the training data
    smote = SMOTE(random_state=42)
    # Reshape for SMOTE if necessary
    X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)

    # Train the model
    history = model.fit(
        X_train_smote, y_train_smote,
        epochs = 10,
        validation_data=(X_test, y_test)
    )


    history_list.append(history)
    stop_epochs.append(len(history.history['loss']))

    # Predict and compute the confusion matrix
    y_pred = model.predict(X_test)
    y_pred_labels = (y_pred > 0.5).astype(int)

    conf_matrix = confusion_matrix(y_test, y_pred_labels)
    plt.figure(figsize=(10, 7))
    sns.heatmap(conf_matrix, annot=True, fmt="d", cmap='Blues',
annot_kws={"size": 16})
    plt.show()
    conf_matrices.append(conf_matrix)

# Calculate the average confusion matrix
conf matrix avg = np.mean(conf matrices, axis=0)
plt.figure(figsize=(10, 7))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap='Blues',
annot kws={"size": 16})
plt.title('Average Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

```
for i, history in enumerate(history_list):
    num_epochs = stop_epochs[i]
    plt.figure(figsize=(12, 5))
    plt.subplot(1, 2, 1)
```

```
    # Ensure 'balanced_accuracy' and 'val_balanced_accuracy' are the
correct keys
    plt.plot(history.history['balanced_accuracy'], label='Train Balanced
Accuracy')
    plt.plot(history.history['val_balanced_accuracy'], label='Validation
Balanced Accuracy')
    plt.title(f'Balanced Accuracy Fold #{i+1}')
    plt.ylabel('Balanced Accuracy')
    plt.xlabel('Epoch')
    plt.axvline(x=num_epochs-1, color='r', linestyle='--')  # Draw a
vertical line
    plt.legend(loc='lower right')

    plt.show()  # Display the figure at the end of each iteration of the
loop
```

```
VN = conf_matrix_avg[0, 0]
FP = conf_matrix_avg[0, 1]
FN = conf_matrix_avg[1, 0]
VP = conf_matrix_avg[1, 1]

# Calcular la sensibilitat i especificitat
sensibilitat = VP / (VP + FN)
especificitat = VN / (VN + FP)

print(f"Sensibilitat: {sensibilitat}")
print(f"Especificitat: {especificitat}")
```