
This is the **published version** of the bachelor thesis:

Setó Llorens, Aina; Casas Roma, Jordi, tut. Anàlisis Multilayer, Multiplex y Aggregated Monoplex de Redes Cerebrales en Esclerosis Múltiple a partir de MRI. 2025. (Enginyeria de Dades)

This version is available at <https://ddd.uab.cat/record/317352>

under the terms of the  license

Anàlisi Multilayer, Multiplex i Aggregated Monoplex de Xarxes Cerebrals en Esclerosi Múltiple a partir de MRI

Aina Setó i Llorens

Resum– Aquest treball se centra en l'anàlisi de les xarxes cerebrals en individus amb esclerosi múltiple utilitzant imatges de ressonància magnètica (MRI). L'objectiu principal consisteix en identificar els patrons d'alteracions en les connexions cerebrals associades amb l'esclerosi múltiple. Per a aquest estudi, s'han utilitzat dades de 270 mostres procedents de dos hospitals situats en països diferents. S'han construït diferents tipus de xarxes per representar la connectivitat cerebral dels individus, i a partir dels embeddings generats de les xarxes s'ha realitzat una detecció de la malaltia mitjançant diferents models de Graph Neural Networks (GNN). Durant el treball projecte també s'han utilitzat tècniques de preprocessament i balanceig de dades per optimitzar l'entrenament dels models, així com també s'ha explorat la influència de l'origen dels conjunts de dades en l'entrenament dels models.

Paraules clau– Esclerosi múltiple, xarxes cerebrals, embeddings, Graph Neural Networks, ressonància magnètica

Abstract– This project focuses on the analysis of brain networks in patients with multiple sclerosis using magnetic resonance imaging (MRI). The main objective is to identify patterns of connectivity alterations in the brain associated with multiple sclerosis. For this study, data from 270 patients from two hospitals located in different countries were used. Different types of networks were constructed to represent the patients brain connectivity, and based on the generated embeddings, disease detection was performed using various Graph Neural Network (GNN) models. During the project, data preprocessing and class balancing techniques were also applied to optimize model training, and the influence of the origin of the datasets on model performance was explored.

Keywords– Multiple sclerosis, brain networks, embeddings, Graph Neural Networks, magnetic resonance imaging

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

L'ESCLEROSI MÚLTIPLE (EM) és una malaltia crònica del sistema nerviós central que afecta el cervell, la medul·la espinal i els nervis òptics que es caracteritza per la destrucció de la mielina, la capa que

protegeix les fibres nervioses [2]. Aquesta malaltia pot afectar persones de totes les edats, tot i que és més comuna en adults joves i en dones. Segons dades de l'Associació Espanyola d'Esclerosi Múltiple (AEDEM-COCEMFE), es calcula que a Espanya hi ha 47.000 persones amb EM, 600.000 a Europa i més de 2.000.000 al món. Tot i que en els darrers anys s'ha observat un augment en la freqüència de la malaltia, incloent-hi casos en nens menors de 15 anys i en adults majors de 65 anys.

L'esclerosi múltiple és una malaltia amb molta variabilitat i es pot manifestar en diverses formes. La síndrome clínic aïllat (CIS) és el primer episodi de símptomes neu-

• E-mail de contacte: aina.seto@autonoma.cat
• Treball tutoritzat per: Jordi Casas Roma (Departament de Ciències de la Computació)
• Curs 2024/25

rològics causats per lesions en la mielina, indicant una possible esclerosi múltiple però sense complir encara els criteris diagnòstics. L'esclerosi múltiple recurrent-remitent (EMRR), que representa el 85% dels casos, es caracteritza per brots on apareixen nous símptomes o empitjoren els existents, seguits de períodes de remissió on aquests poden alleujar-se o desaparèixer. Amb el temps, molts pacients evolucionen cap a l'esclerosi múltiple secundària progressiva (EMSP), on la malaltia progressa de manera contínua sense períodes de remissió. Finalment, l'esclerosi múltiple primària progressiva (EMPP), present en aproximadament un 15% dels casos, es manifesta des del principi amb un empitjorament constant dels símptomes, sense brots diferenciats com en l'EMRR.

La gran variabilitat dins de la malaltia provoca que els símptomes també canvien segons l'estat on es troba el pacient. Els símptomes poden variar àmpliament, però alguns dels més comuns són: problemes de visió on s'inflama el nervi òptic, formigueig a les cames i als braços, fatiga, debilitat muscular i espasmes, problemes d'equilibri i coordinació, dificultat en la concentració, depressió, etc.

El diagnòstic precoç de l'esclerosi múltiple (EM) és fonamental per iniciar tractaments que redueixin les recaigudes i alentir la progressió de la malaltia, millorant així la qualitat de vida dels pacients. Malgrat això, l'esclerosi múltiple no és fàcil de diagnosticar en les seves etapes inicials, ja que no existeixen proves definitives per a la seva detecció. El diagnòstic es basa en l'exclusió d'altres malalties amb símptomes similars mitjançant una combinació de proves, com les imatges de ressonància magnètica (Magnetic Resonance Imaging, MRI) d'entre altres.

Donat que l'esclerosi múltiple no té cura, cal identificar la malaltia com més aviat millor per iniciar tractaments i prevenir la discapacitat. És molt important que es facin estudis de valor sobre aquesta malaltia que permetin un diagnòstic avançat. Per això, aquest treball proposa el desenvolupament de models per a la predicció de la malaltia, així com la investigació de quines parts del cervell són més determinants a l'hora de diagnosticar-la. Gràcies a les imatges de ressonància magnètica, obtenim les xarxes morfològiques, estructurals i funcionals del cervell, i a partir de la combinació d'aquestes tres xarxes com 1 multilayer, multiplex o monoplex que ens permetran tenir una visió més global del funcionament del cervell i comprendre amb més exactitud què passa dins del cervell quan una persona pateix esclerosi múltiple.

2 OBJECTIUS

Els objectius marcats per aquest treball són els següents:

1. **Construir una xarxa cerebral multilayer, multiplex i monoplex:** Integrar les xarxes morfològica, estructural i funcional en una única xarxa que integri la informació de les diferents capes. En un model multiplex, cada node es connecta amb els mateixos nodes de cada capa. En un model multilayer, la xarxa estructural actua com a enllaç entre les altres xarxes, mentre que en un model *aggregated monoplex* les tres capes es fusionen en una sola, sumant els pesos de les connexions de les diferents xarxes.

2. **Definir i aplicar mètriques de teoria de grafs:** Redefinir mètriques de teoria de grafs, com el *strength*, *degree*, *closeness centrality*, entre d'altres, per adaptar-les a grafs amb múltiples capes. Un cop adaptades, aplicar-les per identificar els nodes més importants o rellevants de la xarxa.
3. **Detectar patrons i classificar subjectes mitjançant GNN:** Identificar les regions cerebrals que presenten alteracions en les persones amb esclerosi múltiple i detectar patrons mitjançant models de xarxes neuronals de grafs (GNN).
4. **Analitzar la influència de l'origen del conjunt de dades en els resultats d'un model:** Avaluar si l'origen de les dades afecta el rendiment dels models de detecció d'esclerosi múltiples. Al tenir 2 conjunts de dades de 2 hospitals, es vol comprovar si tenen una distribució similar i si es pot fer servir un mateix model per a tots els individus.

3 METODOLOGIA I PLANIFICACIÓ

La metodologia que s'implementarà durant aquest treball serà la metodologia Cross Industry Standard Process for Data Mining (CRISP-DM). Aquest sistema permet estructurar el projecte en sis fases iteratives de manera organitzada, tal com es mostra a la Figura 1, oferint una gran flexibilitat. Com que és un procés iteratiu, sempre podem tornar enrere si cal fer millores.

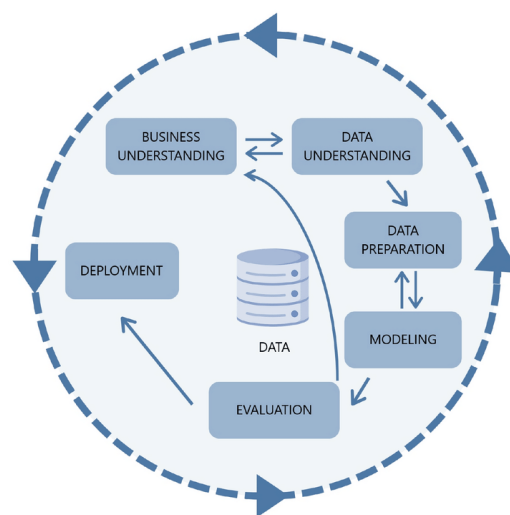


Fig. 1: Fases de la metodologia CRISP-DM [6]

3.1 Comprensió del problema

Aquesta primera fase se centra en el descobriment de la problemàtica i la definició de l'objectiu principal del projecte, en aquest cas analitzar les xarxes cerebrals de persones amb esclerosi múltiple mitjançant tècniques de deep learning. Entre les tasques d'aquesta fase s'inclouen la documentació i investigació sobre la malaltia així com d'estudis prèvis, la definició d'objectius més específics, etc.

3.2 Comprensió de les dades

La segona fase es tracta d'entendre les dades que s'utilitzaran durant el projecte. En aquest cas, les dades provenen d'imatges de ressonància magnètica (MRI) juntament amb fitxers csv etiquetant les imatges.

3.3 Preparació de les dades

L'objectiu d'aquesta fase és la transformació de les dades per ser tractades i analitzades a les fases posteriors. Les tasques pròpies d'aquesta fase seran la construcció de la xarxa multilayer, multiplex i monoplex, neteja i filtratge dels grafs, d'entre altres.

3.4 Modelatge

La fase de modelatge consisteix en la creació de models de deep learning i machine learning per a l'anàlisi de les xarxes. En aquest projecte, s'utilitzaran xarxes neuronals de grafs (GNN) per detectar patrons de deteriorament cerebral i fer una predicció.

3.5 Avaluació

En aquesta fase s'avaluarà el rendiment dels models anteriorment creats per tal d'assegurar que els resultats són fiables. També es farà una comparativa entre els models per veure quin fa una millor classificació de les mostres.

3.6 Interpretació de resultats

L'última fase del cicle consisteix en la interpretació dels resultats obtinguts anteriorment així com extreure conclusions i documentar-ho a l'informe. A més, s'analitzaran quines són les parts del cervell que influeixen més en la predicció de la malaltia, identificant les regions cerebrals clau que poden ser indicatives de deteriorament en pacients amb esclerosi múltiple.

3.7 Planificació

A la Figura 15 de l'apèndix A es presenta una planificació més específica, on els objectius concrets estan associats a un temps determinat.

4 DESENVOLUPAMENT

El codi implementat durant aquest projecte es troba disponible a aquest GitHub: <https://github.com/ainaseto/TFG>. Les dades utilitzades no són públiques per motius de privacitat.

4.1 Anàlisi del conjunt de dades

En aquest estudi s'han utilitzat dades de 270 individus provinents de dos centres hospitalaris diferents. El conjunt 1 està format per 165 persones, de les quals 147 corresponen a la classe 1 (amb esclerosi múltiple) i 18 a la classe 0 (controls sans). El conjunt 2 inclou 105 individus, amb una distribució més equilibrada: 52 positius i 53 negatius. Originalment, les dades dels individus positius hi havia 3 escenaris diferents segons el grau de la malaltia en el que

es trobava, però s'ha simplificat a una classificació binària degut a la poca quantitat de dades de cada tipus. De cada mostra es disposa de tres fitxers csv que contenen les matrius d'adjacència corresponents a tres aspectes diferents de la imatge de la ressonància magnètica: GM (gray matter), RS (resting-state fMRI) i FA (fractional anisotropy). Cada matriu representa les connexions entre diferents regions cerebrals, i permet construir grafs per a l'anàlisi de la seva connectivitat. A més, s'acompanya d'un fitxer de dades demogràfiques i clíniques amb atributs com l'edat, el sexe o l'origen.

Donat que disposem 2 conjunts de dades d'origens diferents, necessitem saber quina és la distribució de les dades de cada conjunt per entrenar un model generalitzat. Les Figures 2 i Figura 3 mostren la distribució del grau mitjà per a les matrius GM, RS i FA, del conjunt 1 tant pels individus negatius com pels individus sans respectivament. A la Figura 2, corresponent als sans, s'observa una distribució de graus entre els valors 43 i 69 per la matriu GM (blau), un pic molt marcat al voltant de 75 per RS (verd), indicant així poca variabilitat de grau. I una distribució més dispersa per FA (vermell) entre 50 i 55. A la Figura 3, que mostra els individus amb esclerosi múltiple, la GM es desplaça lleugerament cap a valors inferiors, mentre que la RS manté el mateix pic al voltant de 75. La FA, tot i mantenir una forma dispersa, presenta valors lleugerament més baixos. Aquestes distribucions ens poden suggerir que GM i FA podrien tenir un paper més clau que RS a l'hora de predir per aquest conjunt.

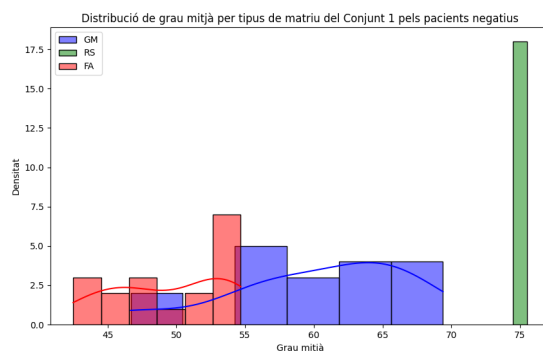


Fig. 2: Distribució de grau mitjà per tipus de matriu del Conjunt 1 pels individus negatius

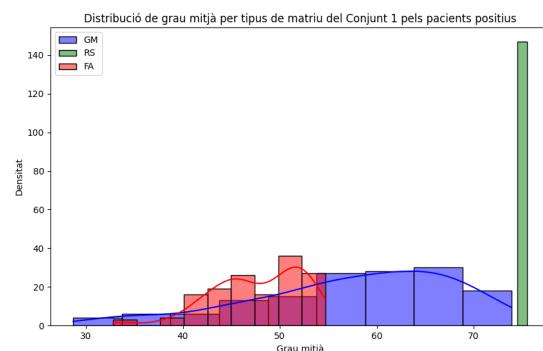


Fig. 3: Distribució de grau mitjà per tipus de matriu del Conjunt 1 pels individus positius

Pel que fa al conjunt 2, tal i com es poden veure a les Figura 4 i Figura 5 les distribucions del grau mitjà mostren diferències clares entre individus positius i negatius, especialment en les matrius GM i FA. La matriu RS és pràcticament idèntica en les dues figures, molt concentrada al valor 75. En canvi, FA mostra una distribució més àmplia en els positius i més estreta en els negatius. En els negatius, GM presenta una major variabilitat respecte les mostres positius on la distribució és entre 63 i 75, indicant una major connectivitat mitjana, mentre que en els negatius és més dispersa, amb valors més baixos.

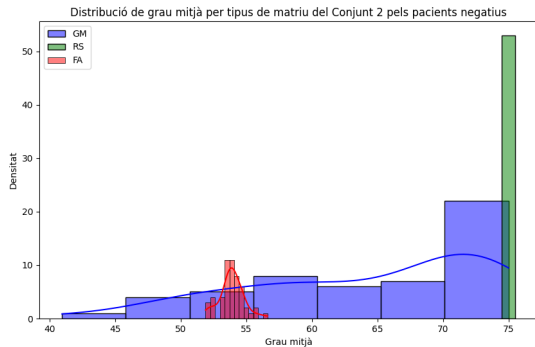


Fig. 4: Distribució de grau mitjà per tipus de matriu del Conjunt 2 pels individus negatius

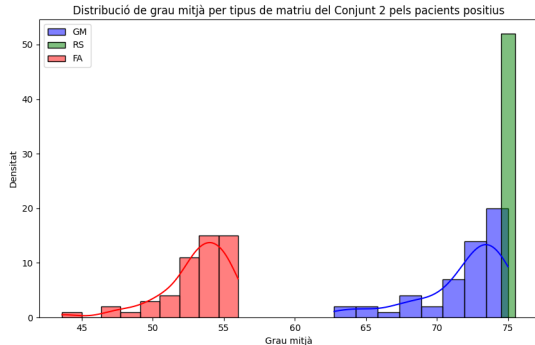


Fig. 5: Distribució de grau mitjà per tipus de matriu del Conjunt 2 pels individus positius

Les dades de la taula 1 mostren les mitjanes i desviacions estàndard dels pesos de les arestes per a tres tipus de matriu (GM, RS i FA) en els dos conjunts, separats segons mostres positius i negatius. Observem que, en general, les diferències entre positius i negatius dins d'un mateix conjunt són petites, especialment en el Conjunt 1, on les mitjanes de GM i RS són molt similars entre grups. En canvi, en el Conjunt 2 hi ha una lleugera tendència a valors més alts de GM i FA en els negatius, mentre que RS es manté bastant constant. També cal destacar que la variabilitat és força similar, amb lleugeres diferències per tipus de matriu.

Aquestes petites diferències entre conjunts podrien tenir un impacte important en un model generalitzat pels 2 conjunts. Si el model detecta patrons específics dels conjunts podria aprendre a detectar entre el conjunt 1 i el conjunt 2 en lloc de detectar la presència de la malaltia.

4.2 Preprocessament de dades

Les xarxes utilitzades per a la detecció d'anomalies al cervell consten de 3 capes, amb 76 nodes cadascuna, que representen diferents parts del cervell. La capa RS (rs-fMRI) representa la correlació funcional entre regions del cervell en estat de repòs, la capa GM significa l'organització de la matèria gris al cervell, que ajuda en el processament de la informació, i finalment, la capa FA representa la connectivitat estructural de la matèria blanca.

Inicialment, les tres capes estaven molt connectades. Les capes RS i GM presentaven 2850 arestes cadascuna, és a dir, el 100% de les arestes possibles en un graf no dirigit ($\frac{76 \times 75}{2}$), amb pesos que oscil·laven entre 0 i 0,8 en el cas de la capa RS (Figura 6.A) i entre 0 i 0,6 en el cas de la capa GM (Figura 6.B). En canvi, a la capa FA, la mitjana d'arestes representava aproximadament el $95,29\% \pm 2,34\%$ respecte al màxim d'arestes possibles, amb pesos que variaven entre 0 i 0,5, i una elevada concentració d'arestes sense pes (Figura 6.C).

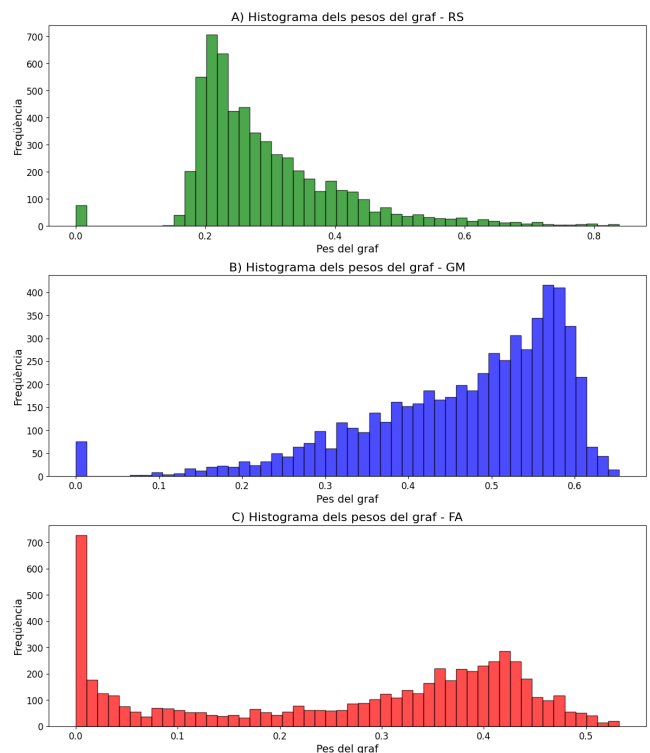


Fig. 6: Histograma dels pesos de les capes RS, GM i FA

Per poder detectar patrons i facilitar el processament de les dades, és necessari fer una neteja del 90% de les arestes i quedar-nos amb les més significatives. En el cas de la capa funcional (RS) i de matèria gris (GM), basant-nos en l'histograma, s'ha aplicat un llindar de 0,44 i 0,59, respectivament, per seleccionar les arestes amb un pes major o igual. Això ha resultat en un volum d'arestes de 576 i 598, que representen un 10,10% i un 10,49% de les arestes originals.

D'altra banda, en la capa estructural (FA), ens interessaven només les arestes que connecten regions cerebrals contigües, per això s'ha utilitzat l'atlas cerebral Desikan-Killiany-Tourville (DKT) [8], que divideix el cervell en 76 regions. La neteja d'aquesta capa ha fet que finalment tinguem 646 arestes, que representen un 11,89%

		Conjunt 1			Conjunt 2		
		GM	RS	FA	GM	RS	FA
Positius	Pes Mitjà	0,5737 ± 0,0228	0,3106 ± 0,2130	0,4106 ± 0,0719	0,5884 ± 0,0267	0,2564 ± 0,1921	0,412 ± 0,0707
	Pes Mínim	0,5397	0,0000	0,0918	0,5383	0,0000	0,1295
	Pes Màxim	0,7508	0,9788	0,6981	0,7318	0,9732	0,6323
Negatius	Pes Mitjà	0,5718 ± 0,0216	0,3063 ± 0,2172	0,4383 ± 0,0783	0,5787 ± 0,0253	0,2609 ± 0,1952	0,4375 ± 0,0728
	Pes Mínim	0,5397	0,0000	0,1408	0,5384	0,0000	0,1176
	Pes Màxim	0,7079	0,9847	0,7038	0,7330	0,9773	0,6812

TAULA 1: ANÀLISI DELS PESOS DE LES ARESTES DELS CONJUNTS DE DADES

de les arestes originals.

La neteja de les arestes es pot veure reflectida en les següents figures. En la Figura 7 i la Figura 8, que corresponen a les capes originals, es pot observar una gran densitat d'arestes, la qual cosa indica una xarxa amb una gran quantitat de connexions. En canvi, a la Figura 9 i la Figura 10, després d'aplicar les màscares anteriorment creades, s'observa una clara disminució en el nombre d'arestes, destacant les més significatives. D'altra banda, a simple vista, no es poden identificar diferències evidents entre les persones sanes i les malaltes.

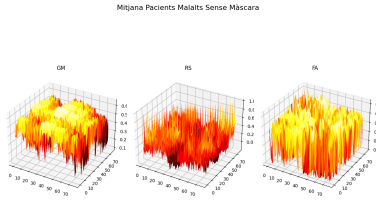


Fig. 7: Mitjana de les capes originals dels individus malalts

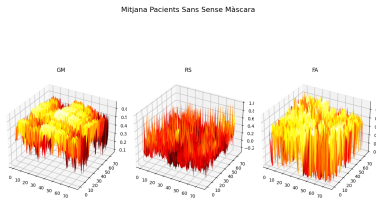


Fig. 8: Mitjana de les capes originals dels individus sans

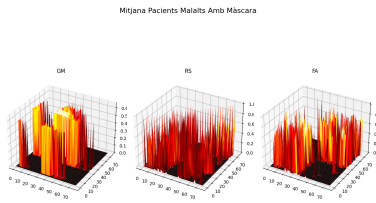


Fig. 9: Mitjana de les capes dels individus malalts aplicant la màscara

Tal com es mostra a la taula 2, després d'aplicar la màscara conjunta, el pes mitjà de les arestes augmenta en

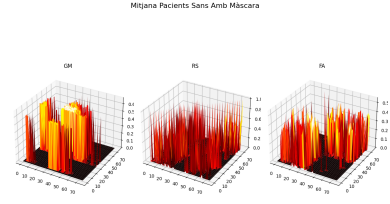


Fig. 10: Mitjana de les capes dels individus sans aplicant la màscara

general, ja que només es conserven les connexions més significatives. Tot i això, en el cas de la capa FA, el pes mitjà disminueix lleugerament. Això es deu al fet que, en aquesta capa, l'objectiu no és prioritzar els pesos més alts, sinó assegurar que les connexions seleccionades siguin de regions pròximes

4.3 Construcció de les xarxes

Per obtenir una estructura de dades que integri els diferents tipus d'informació, cal crear una xarxa que combini les tres capes. En aquest cas, es consideraran tres tipus de xarxes: multilayer, multiplex i aggregated monoplex, cadascuna utilitzant una forma diferent d'enllaç d'informació entre les capes.

4.3.1 Multilayer

Per a la construcció de la xarxa multilayer [1] s'utilitza la capa estructural (FA) com a enllaç entre les capes de matèria gris (GM) i funcional (RS), donant com a resultat una matriu de $(2N) \times (2N)$ on N és el nombre de nodes, en aquest cas 76. La matriu de supra-adjacència de la xarxa ($M_{i,j}^{a,b}$) es defineix segons la Figura 11, i segueix la següent fórmula:

$$M_{i,j}^{1,1} = A_{i,j}^{(GM)} \quad \forall i, j \in \{1, \dots, N\}$$

$$M_{i,j}^{2,2} = A_{i,j}^{(RS)} \quad \forall i, j \in \{1, \dots, N\}$$

$$M_{i,j}^{1,2} = M_{i,j}^{2,1} = A_{i,j}^{(FA)} \quad \forall i, j \in \{1, \dots, N\}$$

4.3.2 Multiplex

La xarxa multiplex [7] connecta tots els nodes de les mateixes capes amb els mateixos nodes de cada capa, donant com a resultat una matriu de $(3N) \times (3N)$ on N és el nombre

		Conjunt 1			Conjunt 2		
Positiu		GM	RS	FA	GM	RS	FA
	Pes Mitjà	0,6169 ± 0,0257	0,4929 ± 0,2229	0,339 ± 0,5600	0,6018 ± 0,0194	0,4511 ± 0,2119	0,4041 ± 0,0763
Negatiu							
	Pes Mitjà	0,6411 ± 0,0212	0,4067 ± 0,2297	0,3399 ± 0,0509	0,6385 ± 0,0167	0,4907 ± 0,2444	0,3457 ± 0,0634

TAULA 2: ANÀLISI DELS PESOS DE LES ARESTES DELS CONJUNTS DE DADES DESPRÉS D'HAVER APLICAT UNA MÀSCARA CONJUNTA

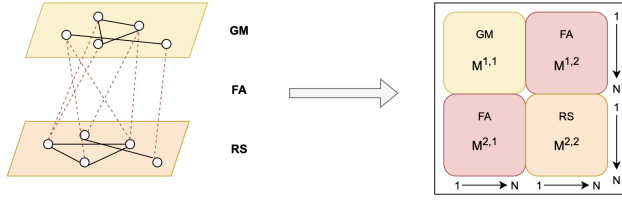


Fig. 11: Esquema de la matriu de supra-adjacència de la xarxa Multilayer

de nodes, en aquest cas 76. Els blocs diagonals de la matriu corresponen a les matrius d'adjacència convencionals dins de cada capa individual, mentre que la connectivitat entre capes es reflecteix en els blocs situats fora de la diagonal. La matriu de supra-adjacència de la xarxa ($M_{i,j}^{a,b}$) es defineix segons la Figura 12, i segueix la següent fórmula:

$$\begin{aligned}
 M_{i,j}^{1,1} &= A_{i,j}^{(GM)} \quad \forall i, j \in \{1, \dots, N\} \\
 M_{i,j}^{2,2} &= A_{i,j}^{(RS)} \quad \forall i, j \in \{1, \dots, N\} \\
 M_{i,j}^{3,3} &= A_{i,j}^{(FA)} \quad \forall i, j \in \{1, \dots, N\} \\
 M_{i,j}^{1,2} &= A_{i,j}^{1,3} = A_{i,j}^{2,1} = A_{i,j}^{2,3} = A_{i,j}^{3,1} = A_{i,j}^{3,2} = I_{76}
 \end{aligned}$$

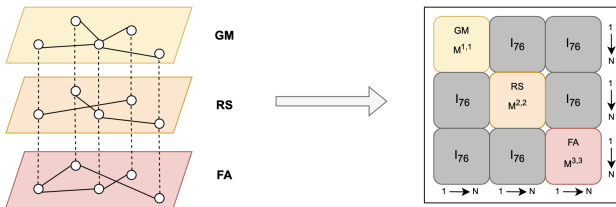


Fig. 12: Esquema de la matriu de supra-adjacència de la xarxa Multiplex

4.3.3 Aggregated Monoplex

La xarxa aggregated monoplex [7] és la xarxa més simple de les 3 on es sumen tots els pesos de les arestes que corresponen a cada capa i es normalitza entre els valors 0 i 1, com les altres xarxes, donant com a resultat una matriu de $(N) \times (N)$ on N és 76. La matriu d'adjacència de la xarxa (M) es defineix segons la Figura 13, i segueix la següent fórmula:

$$M = \sum_{k=1}^3 A_k$$

On:

$$\begin{aligned}
 A_1 &= A_{i,j}^{(GM)} \quad \forall i, j \in \{1, \dots, N\} \\
 A_2 &= A_{i,j}^{(RS)} \quad \forall i, j \in \{1, \dots, N\}
 \end{aligned}$$

$$A_3 = A_{i,j}^{(FA)} \quad \forall i, j \in \{1, \dots, N\}$$

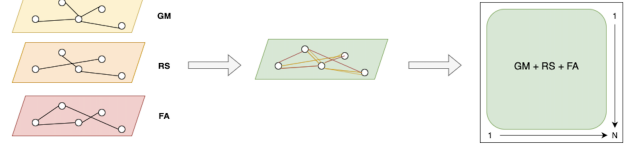


Fig. 13: Esquema de la matriu d'adjacència de la xarxa Aggregated Monoplex

4.4 Creació d'embeddings

Com que no es disposem d'informació prèvia sobre els nodes, ha estat necessari construir embeddings inicials a partir de les pròpies xarxes, que s'utilitzaran posteriorment a la GNN com a informació a propagar. Per generar-los, s'han utilitzat diverses mètriques de teoria de grafs adaptades a les tres xarxes. Aquestes mètriques permeten identificar la importància de cada node dins de cada xarxa. Les mètriques s'han calculat de manera individual per a cada subjecte i per a cada capa, i posteriorment s'han concatenat per formar un vector representatiu únic per a cada node. Les mètriques utilitzades són les següents:

4.4.1 Degree

El grau d'un node indica amb quants altres nodes està connectat. Un alt grau indica que un vèrtex té moltes connexions amb altres vèrtexs del graf i, per tant, pot tenir més influència dins la xarxa. La seva fórmula és:

$$d_i = \sum_{\alpha, \beta, j} \Theta(M_{i,j}^{\alpha, \beta})$$

on Θ és una funció binaritzadora, que retorna 1 si $M_{i,j}^{\alpha, \beta} > 0$ i 0 en cas contrari.

4.4.2 Strength

La força d'un node es defineix com la suma dels pesos de les arestes que té. La força total d'un node es calcula sumant les seves forces en cadascuna de les capes. La seva fórmula és:

$$s_i = \sum_{\alpha} s_i^{\alpha}$$

On:

$$s_i^{\alpha} = \sum_{\beta, j} M_{i,j}^{\alpha, \beta}$$

4.4.3 Clustering

El coeficient de clustering mesura com de connectats estan els veïns d'un node entre ells. La seva fórmula és:

$$c_u = \frac{1}{\deg(u)(\deg(u) - 1)} \sum_{vw} (\hat{w}_{uv} \hat{w}_{uw} \hat{w}_{vw})^{1/3}$$

On:

$$\hat{w}_{uv} = \frac{w_{uv}}{\max(w)}$$

és el pes normalitzat de l'aresta entre u i v , i $\deg(u)$ és el grau del node u .

4.4.4 Closeness Centrality

Aquesta mètrica mesura com de proper és un node a tots els altres. Representa la longitud mitjana del camí més curt entre el node objectiu i tots els altres nodes. Un node amb una alta centralitat de proximitat pot arribar ràpidament a tots els altres vèrtexs de la xarxa. La seva fórmula és:

$$c_i = \left(\sum_{j \neq i} D_{i,j} \right)^{-1}$$

on $D_{i,j}$ representa la distància entre el node i i el node j .

4.4.5 Betweenness Centrality

Aquesta centralitat mesura el nombre de vegades que un node es troba en el camí més curt entre parelles d'altres nodes del graf. La seva fórmula és:

$$b_i = \frac{1}{n(n-1)} \sum_{o,d} \frac{\sigma_{o,d}(i)}{\sigma_{o,d}}$$

on $\sigma_{o,d}$ és el nombre total de camins més curts entre els nodes o i d , i $\sigma_{o,d}(i)$ és el nombre de camins més curts que passen pel node i .

4.4.6 Local Efficiency

L'eficiència local d'un node mesura com de ben connectats estan els seus veïns entre ells quan aquest node és eliminat. La seva fórmula és:

$$LE_i = \frac{1}{d_i(d_i - 1)} \sum_{j,k \in G_i} \frac{1}{D_{j,k}}$$

on d_i és el grau del node i , G_i és el subgraf format pels veïns de i , i $D_{j,k}$ és la distància més curta entre els nodes j i k dins d'aquest subgraf.

4.4.7 PageRank

La centralitat PageRank mesura la importància d'un node tenint en compte no només el nombre d'arestes, sinó també la importància d'aquests. La seva fórmula és:

$$p_i = \frac{1-d}{N} + d \sum_j \frac{w_{ji}}{s_j^{\text{out}}} p_j$$

4.5 Graph Neural Networks

Per a la detecció d'esclerosi múltiple i la classificació de les xarxes, s'utilitzen models de deep learning en grafs, coneguts com Graph Neural Networks (GNN) [9]. Aquests models estan dissenyats per treballar amb dades que tenen una estructura de grafs, cosa que ens permet capturar les relacions entre les regions cerebrals i classificar el graf en funció de la quantitat d'informació que es transmet entre elles. Els embeddings que hem creat anteriorment, el vector amb informació per a cada node, serveixen com a dades d'entrada per al model juntament amb la matriu d'adjacència de cada xarxa. Mitjançant un procés iteratiu en què els missatges entre nodes s'agreguen, el model aprèn els patrons associats a la presència o no d'esclerosi múltiple. S'han provat 3 tipus de GNN: Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN) i Graph Attention Network (GAT).

4.5.1 Graph Convolutional Network (GCN)

La Graph Convolutional Network (GCN) és l'arquitectura que serveix com a base per entendre com és una GNN i va ser introduïda per Kipf i Welling l'any 2017 [10]. L'arquitectura de la GCN utilitzada consta d'una capa convolucional, que transforma els embeddings en un vector de mida 32, seguit d'una funció d'activació ReLU. Després s'aplica una agregació global mitjançant la mitjana per obtenir una representació única per a cada graf. Aquesta representació es passa per una capa de *dropout* del 0,5 per prevenir el sobreentrenament. I finalment, s'aplica a aquest una capa lineal (*fully connected*) on transforma el vector de mida 32 a un vector de 2 valors corresponent a les dues classes, on la funció *softmax* ens dirà quina serà la probabilitat de que pertanyi a cada classe.

S'utilitza només una capa per evitar el "Graph Expansion Problem" [11], ja que en grafs densament connectats i amb models amb moltes capes, els nodes poden acabar compartint la mateixa informació, la qual cosa fa que es perdi la capacitat de diferenciació entre ells.

4.5.2 Graph Isomorphism Network (GIN)

La Graph Isomorphism Network (GIN) va ser introduïda per Keyulu Xu al 2019 [12]. La principal diferència respecte a les altres arquitectures és l'ús de funcions d'agregació i combinació. La funció d'agregació selecciona els veïns i recull la informació d'aquests, mentre que la funció de combinació integra la informació dels veïns per tal de crear un nou vector. Les funcions han de ser injectives per tal d'evitar que es generin sempre els mateixos embeddings. Per aconseguir-ho, s'implementen mitjançant una MLP de mínim dues capes, tal com recomana l'autor.

L'arquitectura consta de dues capes GIN on cada una té una MLP amb dues capes, normalització per batch i funcions d'activació ReLU. La primera capa transforma el embedding de 7 a 32 mentre que la segona capa manté aquesta dimensió. Un cop fetes les capes convolucionals es fa una agregació global amb *global.add.pool* per tal d'ajuntar la informació de tots els nodes en un mateix vector per seguidament concatenar els 2 vectors obtinguts, aquest es passa per dues capes lineals amb ReLU i *dropout*. Finalment, s'aplica una funció *softmax* per determinar la classificació del

subjecte.

4.5.3 Graph Attention Network (GAT)

La Graph Attention Network (GAN) va ser introduïda per Petar Veličković al 2018 [13] i va néixer de la idea que uns nodes són més importants que altres, a diferència de les GCNs, que fan una mitjana o suma uniforme dels veïns, les GAT aprenen quins veïns són més rellevants detectant millor els patrons encara que hi hagi soroll. L'arquitectura consta de dues capes d'atenció. Es fan operacions de *dropout* abans i després de cada capa per evitar el sobreentrenament, i es fa servir la funció d'activació ELU (Exponential Linear Unit). Després de les capes d'atenció, s'agrega la informació dels nodes mitjançant *global_mean_pool*, calculant la mitjana dels embeddings per obtenir una representació del graf. Aquesta es passa per una capa lineal i una funció *softmax* per obtenir la predicció.

4.5.4 Preprocessament, balanceig de classes i entrenament

Un dels problemes a l'hora d'entrenar els models és el desbalanceig de les classes, ja que en un principi hi havia 71 controls sans i 199 persones amb esclerosi múltiple. És per això que, per tal que el model aprengui bé els patrons, cal augmentar el número de mostres sanes. Per millorar el rendiment dels models, s'ha aplicat la tècnica SMOTE (Synthetic Minority Over-sampling Technique) [14] per generar 128 nous grafs de la classe 0. Aquesta tècnica permet crear noves mostres sintètiques a partir de les xarxes, s'agafen 2 persones sanes i després es generen noves mostres sintètiques creant punts entre els 2 individus, tal com es veu a la figura Figura 14.

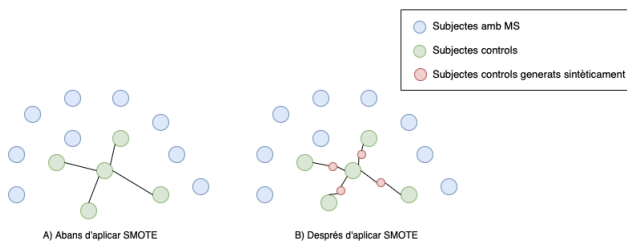


Fig. 14: Esquema de la tècnica SMOTE

Per altra banda, per assegurar un balanceig de classes en cada partició s'ha utilitzat la classe *StratifiedShuffleSplit* de la llibreria *scikit-learn* per separar les dades en conjunts d'entrenament, validació i test. Aquesta estratègia assegura que la proporció de classes es manté constant a cada subconjunt, millorant el rendiment del model.

Per a l'entrenament dels models GNN s'ha utilitzat la funció de pèrdua *CrossEntropyLoss* juntament amb l'optimitzador *Adam*. El *learning rate* s'ha fixat a 0,001 i tots els models s'han entrenat durant 100 èpoques. Les mètriques utilitzades per avaluar els models han estat l'*accuracy*, que mesura la proporció de prediccions correctes en relació al total de prediccions realitzades, i el *recall* tant positiu com negatiu. Aquestes últimes mètriques s'han escollit per analitzar la fiabilitat del model a l'hora de predir si un subjecte té o no té la malaltia.

4.6 Resultats

Després de la construcció dels models, s'han realitzat diversos experiments per avaluar com bons són els models de xarxes neuronals de grafs (GNN) predint si una persona té o no esclerosi múltiple. Amb l'objectiu de veure quanta influència té l'origen de les dades i la seva distribució en el rendiment, s'han plantejat 3 experiments:

1. **Avaluació global del conjunt de dades complet:** S'ha utilitzat el conjunt de dades complet (270 mostres), combinant mostres del conjunt 1 i el conjunt 2.
2. **Avaluació separada per origen:** S'ha entrenat els models per separat utilitzant únicament les persones del conjunt 1 (165 mostres) o del conjunt 2 (106 mostres).
3. **Avaluació separada per origen amb dades sintètiques:** Degut a la poca quantitat de mostres del conjunt 2, s'ha generat noves dades sintètiques.

Per a cada experiment s'ha utilitzat les arquitectures de GNN presentades anteriorment (GCN, GIN i GAT), i s'ha avaluat tenint en compte 3 mètriques detallades prèviament.

4.6.1 Avaluació global del conjunt de dades complet

El primer experiment consisteix en la creació de tres models per a cada xarxa creada anteriorment. El conjunt utilitzat està format per 270 mostres, provinents tant del conjunt 1 com del conjunt 2. D'aquests individus, 70 eren de control i 199 tenien esclerosi múltiple, però per tal de balancejar les classes i garantir que el model sigui capaç de detectar amb la mateixa eficàcia les dues classes, es van generar dades sintètiques per a la classe de mostres sanes fins a obtenir 199 mostres. L'objectiu d'aquest primer experiment és avaluar si un model generalitzat per als dos conjunts pot donar bons resultats en la classificació d'individus.

Model	Mètrica	Multilayer	Multiplex	Monoplex
GCN	Accuracy	50.78%	59.38%	50.00%
	Recall pos.	3.12%	23.44%	100.00%
	Recall neg.	98.44%	95.31%	0.00%
GIN	Accuracy	69.53%	72.66%	67.97%
	Recall pos.	90.60%	84.76%	86.75%
	Recall neg.	34.23%	71.62%	25.68%
GAT	Accuracy	64.84%	68.75%	40.62%
	Recall pos.	76.88%	74.29%	79.64%
	Recall neg.	53.90%	63.95%	4.84%

TAULA 3: RESULTATS DELS MODELS GCN, GIN I GAT PER A DIFERENTS TIPUS DE XARXA PER EL CONJUNT DE DADES COMPLET.

Tal com es mostra a la taula 3, el model de GNN que ens dona més bons resultats en les 3 xarxes és la GIN, especialment en la xarxa multiplex. Per contra, amb la GCN obtenim els pitjors resultats, mentre que amb la GAT tenim uns valors més variables. També cal destacar que la xarxa monoplex no millora els resultats en la detecció de la malaltia respecte a les arquitectures multilayer i multiplex, mentre que aquesta última, al ser més complexa, ens dona millors resultats.

4.6.2 Avaluació separada per origen

El segon experiment consisteix en la creació de models separats segons l'origen de les dades, amb l'objectiu de valorar si el rendiment dels models millora al entrenar-los de manera independent per a cada conjunt de dades, o per contra els models generalitzats creats en l'experiment anterior ja són suficients.

Conjunt 1: Per l'experiment amb les dades del conjunt 1, es van utilitzar 165 mostres, dels quals 18 eren de classe 0 i 147 de classe 1. En aquest cas, també es van balancejar les classes generant dades sintètiques per augmentar la classe 0 fins a 147 mostres.

Conjunt 2: En el cas de l'experiment amb el conjunt 2, aquest està format per 105 persones, amb 54 de classe 0 i 52 de classe 1. A diferència del conjunt 1, no s'ha aplicat cap augment de dades, ja que les dues classes estaven pràcticament balancejades.

Tal com es mostra en la taula 4, tots els models de les xarxes multilayer i multiplex, l'accuracy i els recalls són superiors al 70%, mentre que en els models amb la xarxa monoplex els resultats són lleugerament més baixos. Pel que fa als tipus de GNN, no hi ha gaire variància, tot i que la GIN ofereix una mica millors resultats en comparació amb el GCN i el GAT. En aquest cas, la xarxa multilayer amb el model GIN és la que ofereix els millors resultats, amb una accuracy i recalls superior al 80%.

Model	Mètrica	Multilayer	Multiplex	Monoplex
GCN	Accuracy	72.88%	74.58%	66.10%
	Recall pos.	68.97%	72.41%	68.97%
	Recall neg.	76.67%	76.67%	63.33%
GIN	Accuracy	81.36%	77.97%	67.80%
	Recall pos.	82.76%	72.41%	41.38%
	Recall neg.	80.00%	83.33%	93.33%
GAT	Accuracy	76.27%	71.19%	66.10%
	Recall pos.	75.86%	79.31%	44.83%
	Recall neg.	76.76%	63.33%	86.67%

TAULA 4: RESULTATS DELS MODELS GCN, GIN I GAT PER A DIFERENTS TIPUS DE XARXA PER EL CONJUNT DE DADES 1.

Com es pot veure la taula 5, en aquest cas, els resultats d'accuracy són més baixos en general. Els resultats estan al voltant del 50%, fet que podria indicar que les prediccions es realitzen de manera gairebé aleatòria, ja que no s'obtenen millors resultats de manera consistent entre les diferents xarxes i models.

4.6.3 Avaluació separada per origen amb dades sintètiques

El tercer experiment neix com a continuació de l'experiment 2 amb el conjunt 2, on s'observen resultats considerablement pitjors respecte les altres proves. L'objectiu d'aquest és identificar si el rendiment és més baix perquè el model no és capaç de predir correctament sobre aquest conjunt específic de dades, o bé per la poca quantitat de mostres que té. Per tal de provar si està relacionat amb el nombre de mostres, s'ha augmentat les 2 classes amb dades sintètiques fins a obtenir 150 individus de cada classe, i per tant veure si augmentant la mida del conjunt millora el rendiment.

Model	Mètrica	Multilayer	Multiplex	Monoplex
GCN	Accuracy	52.38%	57.14%	52.38%
	Recall pos.	20.00%	30.00%	0.00%
	Recall neg.	81.82%	81.82%	100.00%
GIN	Accuracy	52.38%	52.38%	47.62%
	Recall pos.	0.00%	0.00%	100.00%
	Recall neg.	100.00%	100.00%	0.00%
GAT	Accuracy	57.14%	61.90%	57.14%
	Recall pos.	50.00%	50.00%	40.00%
	Recall neg.	63.64%	72.73%	72.73%

TAULA 5: RESULTATS DELS MODELS GCN, GIN I GAT PER A DIFERENTS TIPUS DE XARXA PER EL CONJUNT DE DADES DE 2.

Els resultats de la taula 6 no disten gaire de l'experiment sense dades sintètiques, amb valors d'accuracy que ronden el 50%, cosa que podria indicar una predicció aleatòria. En molts casos, també es detecta una gran diferència entre el recall de la classe positiva i el de la negativa, indicant desequilibri en la classificació, fins i tot en algun cas el model només encerta una de les dues classes. Cap combinació ens dona uns resultats fiables per a la detecció de la malaltia.

Model	Mètrica	Multilayer	Multiplex	Monoplex
GCN	Accuracy	55.00%	50.00%	50.00%
	Recall pos.	56.67%	20.00%	100.00%
	Recall neg.	53.33%	80.00%	0.00%
GIN	Accuracy	68.33%	48.33%	53.33%
	Recall pos.	50.00%	53.33%	86.67%
	Recall neg.	86.67%	43.33%	20.00%
GAT	Accuracy	55.00%	43.33%	61.67%
	Recall pos.	86.67%	20.00%	33.33%
	Recall neg.	23.33%	66.67%	90.00%

TAULA 6: RESULTATS DELS MODELS GCN, GIN I GAT PER A DIFERENTS TIPUS DE XARXA PER EL CONJUNT DE DADES DE 2 AUGMENTAT AMB DADES SINTÈTIQUES.

4.7 Anàlisi dels resultats

Els resultats obtinguts mostren diferències rellevants en el rendiment dels models segons el tipus de xarxa utilitzada i l'arquitectura de GNN. En general, les xarxes multiplex i multilayer han tingut millors resultats que les xarxes monoplex. Això es deu al fet que les dues primeres xarxes conserven la separació de les capes funcionals, estructurals i morfològiques obtenint una estructura més complexa que ens permet identificar millors patrons. Per contra, la monoplex simplifica massa les capes, al sumar-les en una sola capa i tracta tota la informació per igual. Aquesta fusió fa que es perdin detalls importants i que el model tingui més dificultats per detectar patrons diferenciats.

Pel que fa a les arquitectures de GNN, el GIN ha sigut el que ens ha donat millors resultats en la majoria de casos, ja que té una arquitectura que ens permet detectar millor les relacions entre els nodes i, per tant, una millor classificació. Mentre que el GCN ha mostrat menor eficàcia detectant la malaltia perquè fa una mitjana simple de la informació dels nodes veïns, i això pot provocar una pèrdua d'informació. El GAT, tot i que incorpora un mecanisme d'atenció per

donar més pes a certes connexions, no ha superat el GIN en cap dels experiments.

Finalment, en els resultats també s'ha detectat que el rendiment del model depèn de l'origen de les dades. Quan els models s'entrenen amb els 2 conjunts, la detecció pot ser enganyosa perquè pot ser que el model aprengui a identificar característiques específiques de l'hospital d'origen en lloc de patrons de malaltia. La variabilitat de les dades respecte als hospitals pot ser per la utilització d'escàners diferents o pacients amb diferents característiques clíniques, això fa que els models no siguin gaire generalitzables.

4.8 Conclusions

Aquest treball ha explorat l'anàlisi de xarxes cerebrals multilayer, multiplex i monoplex a partir d'imatges de resonància magnètica amb l'objectiu de detectar l'esclerosi múltiple mitjançant models de Graph Neural Networks (GNN). A través de la combinació de diferents capes (GM, RS i FA), s'ha construït grafs que permeten capturar patrons de la connectivitat del cervell en persones amb la malaltia i de persones sanes. Mitjançant l'ús de mètriques de teoria de grafs s'han generat embeddings dels nodes, que han servit com a entrada per entrenar diversos models de GNN. Els resultats obtinguts mostren que les xarxes multiplex i multilayer, juntament amb el model GIN, són les que ofereixen una millor capacitat de classificació, mentre que les xarxes monoplex han estat menys eficients.

Pel que fa als objectius plantejats inicialment s'han assolit tots. S'han construït les diferents xarxes, s'han adaptat i aplicat mètriques específiques per a grafs amb diverses capes, s'han entrenat models de GNN per detectar la malaltia i s'ha analitzat com l'origen de les dades pot afectar el rendiment dels models. Sobre aquest últim objectiu, s'ha mostrat que hi ha una dependència important entre el rendiment del model i la variabilitat dels conjunts, evidenciant la necessitat de tenir dades harmonitzades per poder utilitzar un model generalitzat.

Tot i haver complert els objectius inicials, el treball també presentava diverses limitacions. D'una banda, el conjunt de dades estava desbalancejat, amb molts més mostres amb esclerosi múltiple que controls sans. A més, la mida reduïda de persones, especialment del segon conjunt de dades i les diferències de la distribució de dades entre hospitals han fet que els models obtinguts no siguin completament generalitzables.

Com a futures millores, seria interessant ampliar la classificació per incloure els diferents tipus d'esclerosi múltiple, més enllà de la detecció binària utilitzada en aquest treball. Aquesta classificació més detallada podria aportar una visió més realista i útil des del punt de vista clínic, ja que permetria no només detectar la presència de la malaltia, sinó també identificar en quin estat o fase es troba la persona. D'altra banda, també seria necessari tractar les diferències entre els dos conjunts de dades. Aconseguir una distribució més uniforme entre conjunts ajudaria a millorar la capacitat de generalització dels models i reduiria el risc que aquests aprenguin patrons propis d'un hospital concret en lloc de patrons reals de la malaltia. També seria interessant explorar l'opció d'incorporar més informació clínica o biomarcadors del pacient per fer una detecció més precisa i poder fer un diagnòstic més precoç i precís.

REFERÈNCIES

- [1] J. Casas-Roma *et al.*; *Applying multilayer analysis to morphological, structural, and functional brain networks to identify relevant dysfunction patterns*. Network Neuroscience, 2022; 6(3): 916–933. doi: https://doi.org/10.1162/netn_a-00258.
- [2] Multiple Sclerosis, World Health Organization (WHO). <https://goo.su/wPQHox>.
- [3] Esclerosi múltiple, Ministerio de Sanidad de España. <https://goo.su/gpaq>.
- [4] Carme Escalés; *La importància del tractament precoç en l'esclerosi múltiple*. Catalunya Plural. <https://goo.su/AaE1Zm>.
- [5] Tipos de esclerosi múltiple: variantes y sus características, Fundació Esclerosi Múltiple. <https://goo.su/ahQ7z>.
- [6] Oliveira, D. *et al.*; *Prediction of COVID-19 diagnosis based on openEHR artefacts*. Scientific Reports, 12, 12549 (2022). doi: <https://doi.org/10.1038/s41598-022-15968-z>.
- [7] David Kainer *et al.*; *RWRtoolkit: multi-omic network analysis using random walks on multiplex networks in any species*. bioRxiv. doi: <https://doi.org/10.1101/2024.07.17.603975>.
- [8] Arno Klein and Jason Tourville; *101 labeled brain images and a consistent human cortical labeling protocol*. Frontiers in Neuroscience. doi: <https://doi.org/10.3389/fnins.2012.00171>.
- [9] Maxime Labonne; *Hands-On Graph Neural Networks Using Python: Practical techniques and architectures for building powerful graph and deep learning apps with PyTorch*. Packt Publishing, 2023.
- [10] Thomas N. Kipf and Max Welling; *Semi-Supervised Classification with Graph Convolutional Networks*. International Conference on Learning Representations. doi: <https://doi.org/10.48550/arXiv.1609.02907>.
- [11] mon J.D. Prince.; *Understanding Deep Learning*. MIT Press. <http://udlbook.com>.
- [12] K. Xu, W. Hu, J. Lesjovek, and S. Jegelka; *How Powerful are Graph Neural Networks?*. International Conference on Learning Representations. doi: <https://doi.org/10.48550/arXiv.1810.00826>.
- [13] P. Veličković *et al.*; *How Powerful are Graph Neural Networks?*. International Conference on Learning Representations. doi: <https://doi.org/10.48550/arXiv.1710.10903>.
- [14] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer; *SMOTE: Synthetic Minority Over-sampling Technique*. doi: <https://doi.org/10.1613/jair.953>.

A DIAGRAMA DE PLANIFICACIÓ

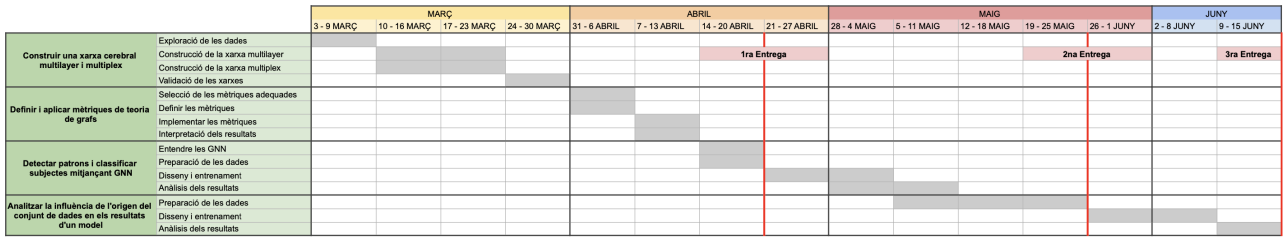


Fig. 15: Diagrama de Gantt de la planificació de tasques