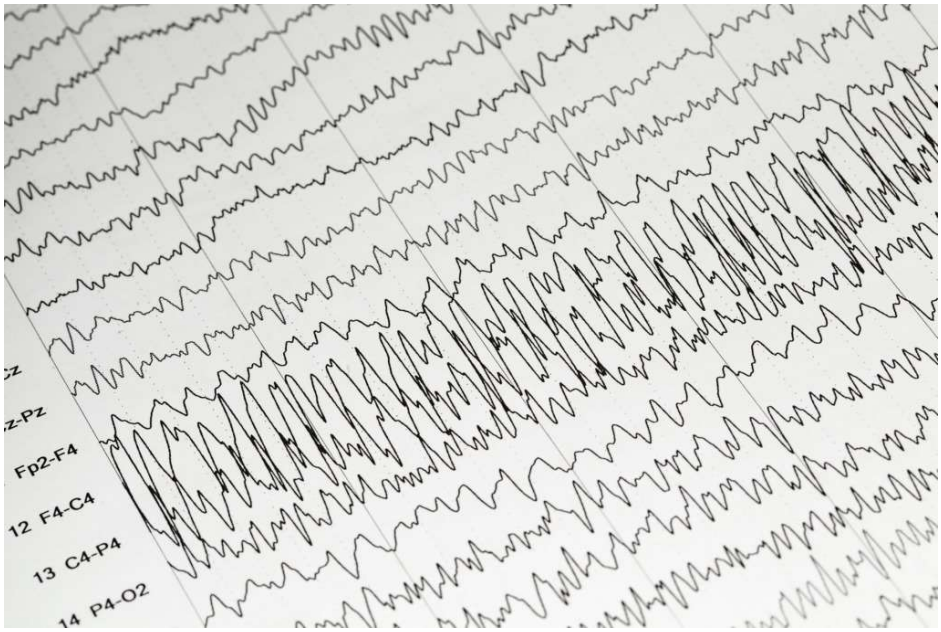
UAB
Universitat Autònoma de Barcelona

Epilèpsia amb electroencefalograma



Submitted by:

Carlota Cortés Mir, 1639080

Paula Macías Alcaide, 1636458

Marc Puigbó Paricio, 1636671

Pol Riubrogent Comas, 1636486

MAPSIV

21/01/2025

1. Introducció

El problema que ens planteja aquest repte es basa en la detecció de convulsions epilèptiques mitjançant l'anàlisi en enregistraments electroencefalogrames, també anomenat (EEG).

Un electroencefalograma és una eina no invasiva que serveix per explorar la funcionalitat del cervell mitjançant l'enregistrament de l'activitat elèctrica de les neurones durant la seva sinapsis.

L'activitat elèctrica de les neurones es rastreja per un conjunt de N elèctrodes col·locats sobre el cuir cabellut del cap. Com es pot observar a la figura 1.

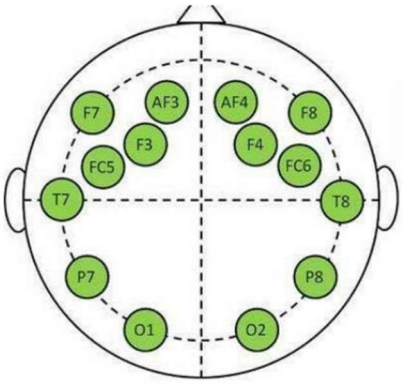


Figura 1. Elèctrodes sobre el cuir cabellut

Els N elèctrodes col·locats en el cuir cabellut proporcionen N senyals temporals 1D (canals), a la figura 2 podem veure un exemple.

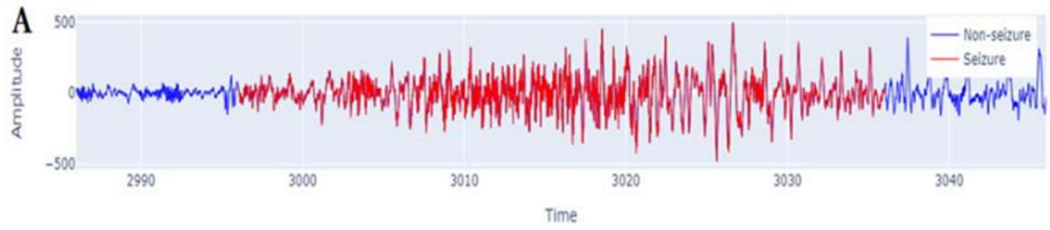


Figura 2. Exemple de senyal EEG anotada

En un procés diagnòstic d'epilèpsia, el neuròleg explora visualment els senyals d'EEG registrats, amb l'objectiu de trobar patrons d'ones afilades, agudes i lentes que caracteritzen una crisi epilèptica.

Aquesta feina pot ser molt lenta i dura, perquè els registres d'EEG poden durar varies hores.

L'objectiu global d'aquest repte és aconseguir classificar aquests patrons d'ones segons si són convulsions o no de forma automàtica.

El link per accedir al git és: https://github.com/NIU1636486/PSIV_Repte4

1.1. Objectius

Per poder aconseguir aquest objectiu global, tindrem en compte altres com:

1- Ús d'arquitectures que ens permetin la fusió de canals EEG.
2- Ús d'arquitectures que tinguin en compte el context temporal.
3- Avaluar el nivell de generalització dels models que implementem.

2. Metodologia

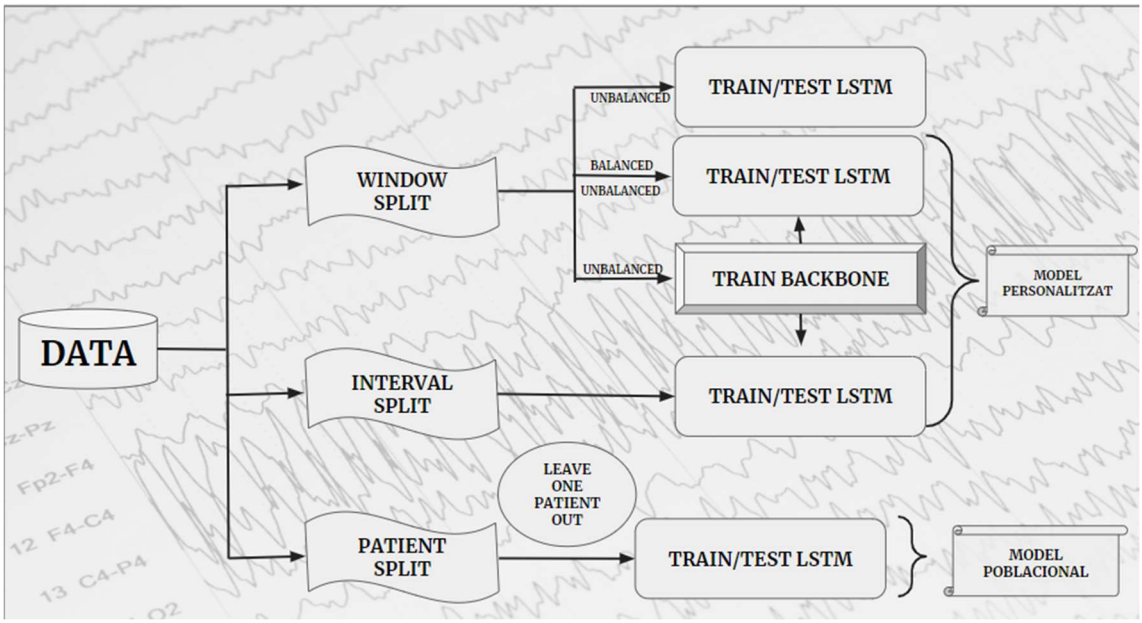


Figura 3. Pipeline del treball

Per tal d'assolir els objectius plantejats, s'ha dissenyat i utilitzat el pipeline de la Figura 3. A partir del dataset preparat proporcionat, hem fet 3 particions:

- Nivell finestra
- Nivell interval
- Nivell pacient

Les particions a nivell finestra i interval ens produeixen models personalitzats, i a nivell de pacient tenim un model poblacional. D'aquesta manera podem avaluar el nivell de generalització amb les diferents particions d'entrenament.

A nivell de finestra s'ha entrenat un model LSTM amb una fusió de canals *flatten*, el model Backbone i finalment una combinació dels 2.

A nivell d'interval (agrupant les dades segons l'episodi global), s'ha entrenat i provat un model LSTM amb backbone.

I per últim a nivell pacient s'ha entrenat i provat el mateix model, però agrupant per pacient, seguint el mètode leave-one-patient-out.

2.1 Arquitectura LSTM

Per tal de complir aquesta tasca, s'ha utilitzat una arquitectura recurrent, més concretament una LSTM. Ens permet no només predir amb les dades d'entrada actuals, sinó que també té en compte el context de la resta de finestres entrades anteriorment. Això afavoreix aquesta tasca concreta, ja que al anar per segments temporals, varies finestres juntes tindran la mateixa classe, i per tant, aquest context temporal ens ajuda a millorar els resultats. S'ha fet servir una arquitectura LSTM simple, amb només una capa, ja que pel que hem trobat, en aquesta tasca concreta era la millor opció.

Com s'ha dit anteriorment, les nostres finestres contenen 21 canals amb 128 mostreigs cada un. Per tal de millorar el rendiment de l'entrenament, s'ha reduït aquesta entrada fusionant els canals. Primer s'ha provat aplanant les dades, així obtenint un sol vector a partir dels 21 canals.

2.2 Arquitectura Backbone

Amb el *flatten*, aconseguim reduir dimensionalment les nostres dades, però el volum segueix sent el mateix. Per millorar-ho, s'ha ideat i entrenat una xarxa neuronal per tal d'extreure característiques de les finestres i així reduir encara més el volum de dades.

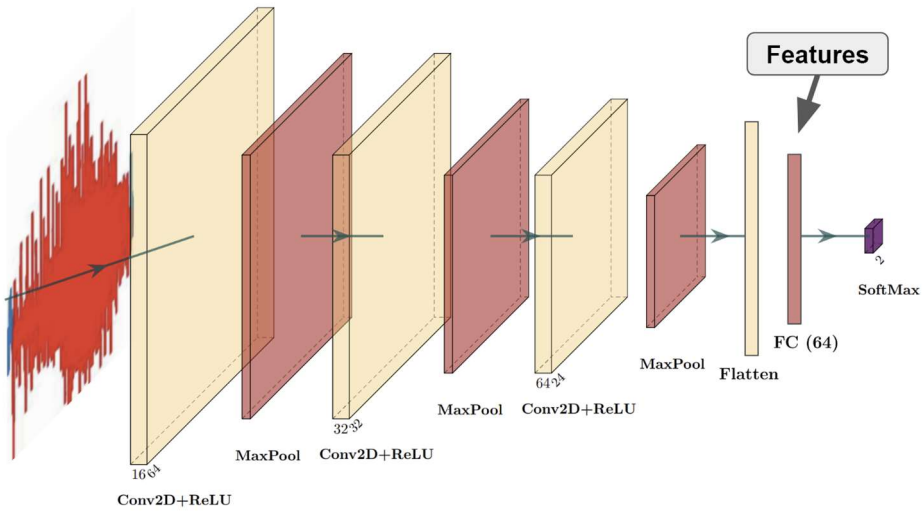


Figura 4. Arquitectura del backbone

Per fer-ho s'ha fet servir una CNN simple [1], amb 3 blocs convolucionals per extreure característiques, i una capa fully-connected que representarà el vector de característiques que posteriorment utilizarem com entrada a la unitat LSTM.

L'arquitectura final és la següent:

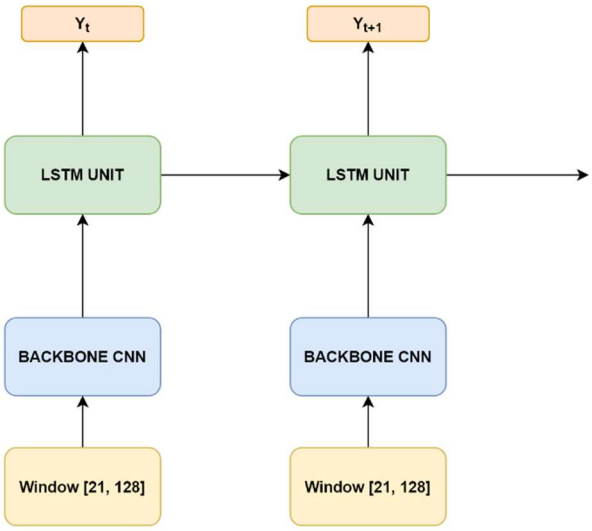


Figura 5. Arquitectura de l'arquitectura completa

3. Disseny Experimental

Per provar aquesta arquitectura hem fet diferents experiments, a nivell de finestra, interval global i pacient. Però primer cal conèixer les dades amb les que treballarem.

3.1 Descripció Dataset

Les dades utilitzades per entrenar i testejar els models implementats provenen del dataset Children's Hospital Boston-MIT.

Aquest dataset conté EEG registrats de 24 pacients amb convulsions severes, els quals són 23 pacients pediàtrics i un adult. Els enregistraments van ser realitzats durant varies hores per pacient i els registres estan desat en fitxers EDF (European Data Format) de aproximadament una hora de duració.

La màquina d'EEG va utilitzar 21 elèctrodes i les senyals van ser recol·lectades amb freqüències de mostreig de 256 Hz.

D'aquestes dades es van processar els registres empaquetats en un parquet, submostrejats a 128 Hz i encapçats en finestres temporals d'1 segon

Aquestes finestres es classifiquen de forma binaria, (1) si formen part d'un episodi convulsional o (0) si són normals, com podem observar a la figura 6

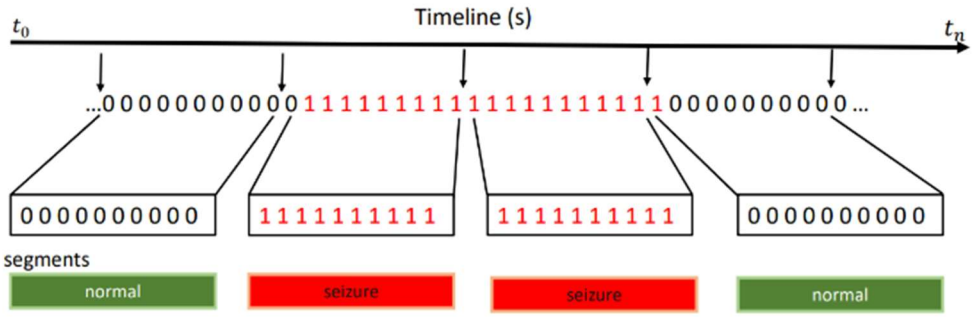


Figura 6. Classificació binaria de les finestres temporals

A la figura 7 podem observar una senyal dividida en diferents intervals que han sigut classificats; en color taronja finestres normals, en vermell finestres on hi ha convulsió i en color groc dos intervals de 30 segons classificats com preictal i postictal. Aquests intervals contenen finestres que presenten patrons de transició; és a dir, els patrons de les finestres no són els mateixos que en els intervals on no hi ha convulsió, ni iguals als intervals on si hi ha, el que dificulta la classificació, i per tant s'han de descartar

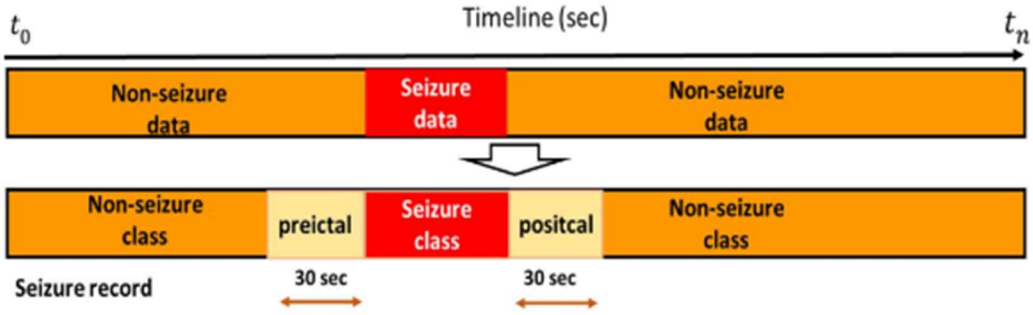


Figura 7. Senyal dividida per intervals

Pels intervals on no hi ha convulsió, les finestres no es superposen però en les finestres on hi ham, tenen un 80% de superposició per fins d'augment de dades.

Totes les finestres temporals (21,128) estan guardades en una matriu tridimensional anomenada EEG_win en un fitxer npz. Sent la primera dimensió el numero de finestres pel pacient.

Per cada pacient, tenim les metadades següents en format parquet:

Class – Variable binaria indicant si hi ha convulsió o no.

Filename_interval – Identifica el interval de l'episodi de la convulsió per cada fitxer .edf

Global_interval – Identifica el interval de l'episodi de la convulsió per pacient

Filename – el nom del registre original edf. És útil per poder entrenar i testejar models personalitzats pel conjunt de pacients.

3.2 Mètriques i Experiments

Experiment 1. Nivell de finestra I

El nivell més baix d'anàlisi correspon al nivell de finestra. En aquest cas, hem dissenyat tres experiments diferents.

En aquest experiment inicial, vam explorar el funcionament de la LSTM passant les dades directament després d'aplicar un flatten. Per fer-ho, vam utilitzar un k-fold amb 5 particions per a entrenament i prova, sense cap agrupació prèvia.

Experiment 2. Nivell de finestra II

Un cop vam verificar que la LSTM funcionava correctament, vam realitzar un segon experiment implementant un backbone per extreure característiques més robustes de les dades.

En aquest experiment, les dades es van utilitzar sense cap tipus de balanceig, fet que ens va permetre analitzar el rendiment de l'arquitectura en un entorn desequilibrat i més proper a la realitat.

Experiment 3. Nivell de finestra III

En el tercer experiment vam abordar el desbalanceig significatiu de les dades i la limitada presència de mostres positives. Vam optar per aplicar un k-fold amb dades balancejades mitjançant submostreig. Aquesta estratègia se centra a reduir el overfitting associat a la classe dominant.

Experiment 4. Nivell de interval

Els experiments següents es van realitzar a nivell d'interval. Per evitar carregar totes les dades a la memòria i prevenir una càrrega excessiva, vam equilibrar-les a nivell de finestra o d'interval, filtrant-les i reduint-ne la quantitat. També vam explorar l'ús de dades no equilibrades, seleccionant un percentatge de finestres o intervals aleatòriament. Després de diversos experiments, vam escollir l'enfocament que va proporcionar els millors resultats.

La LSTM amb backbone es va entrenar utilitzant un GroupKFold amb 5 particions. Els grups es van definir mitjançant la combinació de l'etiqueta, l'interval global i la classe com a claus de grup. Per mitigar el desbalanceig, es va reduir el nombre de finestres, filtrant un percentatge dels grups de classe 0.

Amb l'objectiu de preservar el context temporal i evitar la fuga de dades, vam aplicar una estratègia que assegurava que les finestres d'un mateix interval global no apareguessin tant al conjunt d'entrenament com al de prova. Això és especialment rellevant perquè les crisis epilèptiques són esdeveniments dependents del context temporal de les senyals cerebrals.

Experiment 5. Nivell de pacient

Per a una partició de dades a nivell de pacient, totes les finestres de prova pertanyen exclusivament a un pacient, mentre que les finestres dels altres pacients s'utilitzen per a l'entrenament. Es realitza una divisió k-fold en què, en cada partició, s'exclou un pacient diferent. Aquesta divisió avalua el rendiment del model en un pacient nou mai vist.

En concret, amb un total de 24 pacients, es fan servir les finestres de 23 pacients per entrenar el model i es testen les finestres del pacient restant. Aquest enfocament garanteix que les dades d'entrenament i de prova siguin independents i representa una estratègia més generalista, ja que simula de manera realista el comportament del model amb dades de pacients nous.

En aquest experiment, hem utilitzat dades balancejades i agrupades per pacient amb el model backbone. Durant la càrrega de dades, es va crear la variable group, que identifica els pacients. Això ens va permetre aplicar un k-fold amb 5 particions, testejant un pacient diferent a cada partició.

4. Resultats

4.1 Experiment 1 - Experiment 2

En l'entrenament, els models Backbone i Flatten han mostrat bones mètriques d'Accuracy, F1-score i AUC, amb valors globalment similars entre si. A més de poca desviació estàndard. Els valors dels dos models són aproximadament de 0.89 per Accuracy, 0.73 en F1-Score i 0.86 per AUC, com es pot observar en la figura 8.

Model	Accuracy	F1_score	AUC
Backbone	0.8857 ± 0.0117	0.7267 ± 0.0338	0.8675 ± 0.0205
Flatten	0.8940 ± 0.0046	0.7431 ± 0.0107	0.8624 ± 0.0037

Figura 8. Resultats mean ± std dels models backbone i flatten en l'entrenament NW no balancejat

Tanmateix, no podem extreure molta informació significativa de l'entrenament, ja que en dividir les classes en negatiu (0) i positiu (1) en les mètriques del test es pot observar que el model tendeix a predir preferentment la classe negativa en tots dos models. El que significa valors alts per a negatius i prediccions molt

incorrectes per a positius, amb resultats pitjors que l'atzar (menys de 0.50 en F1-score), com es pot observar en la figura 9.

Model	Recall_0	Recall_1	Precision_0	Precision_1	F1_0	F1_1	Accuracy
Backbone	0.94 ± 0.12	0.25 ± 0.28	0.88 ± 0.06	0.66 ± 0.25	0.90 ± 0.05	0.24 ± 0.18	0.83 ± 0.07
Flatten	0.93 ± 0.02	0.32 ± 0.08	0.89 ± 0.03	0.47 ± 0.10	0.91 ± 0.02	0.38 ± 0.08	0.84 ± 0.03

Figura 9. Resultats mean ± std dels models backbone i flatten en test NW no balancejat

En altres paraules, els resultats mostren una gran variabilitat entre les dues classes. Són molt bons per predir la classe negativa, però fallen en predir la classe positiva.

Com es pot observar en la figura 9, per a negatius, el recall (0.94 ± 0.12/ 0.93 ± 0.02) i la precisió (0.88 ± 0.06/ 0.89 ± 0.03) són molt bons. En canvi, per a positius, el recall (0.25 ± 0.28/ 0.32 ± 0.08) i la precisió (0.66 ± 0.25/ 0.47 ± 0.10) són extremadament inferiors. En general, el model no aconsegueix generalitzar bé per a la classe positiva, donant prioritat a les prediccions de la classe negativa.

Aquest desequilibri entre classes es reflecteix també en les matrius de confusió acumulades dels cinc folds del test. On els true positives són 9.670-16.067, però per true negatives són 250.000 aproximadament, com es pot observar en les figures 10 i 11.

250.022	16.978
37.889	9.670

Figura 10. Matriu Confusió Acumulada Backbone en test NW no balancejat

249.016	17.640
31.836	16.067

Figura 11. Matriu Confusió Acumulada Flatten en test NW no balancejat

4.2 Experiment 3

Els resultats de les mètriques Accuracy, F1-score i AUC en l'entrenament amb dades balancejades continuent sent bones pel model backbone. El F1-score i AUC són similars als obtinguts en l'entrenament amb dades no balancejades. Tanmateix, l'Accuracy ha disminuït, passant de 0.89, en el cas de dades no balancejades, a 0.78 amb dades balancejades, com es pot observar en la figura 12.

	Accuracy	F1-Score	AUC
Mean ± std	0.7804 ± 0.0117	0.7799 ± 0.0117	0.8329 ± 0.0195

Figura 12. Resultats mean ± std del model backbone en l'entrenament NW balancejat

A més, en el test amb dades balancejades s'observen més diferències entre els cinc folds en comparació amb l'experiment amb dades no balancejades. Per exemple, en el fold 1, s'ha obtingut un recall_1 de 0.98, mentre que en el fold 4 el recall_1 ha baixat fins a 0.30, una notable variabilitat entre els folds, com es pot observar en la figura 13.

Tot i això, analitzant la mitjana i desviació estàndard, es pot observar un millor equilibri entre les classes 0 i 1 en comparació amb el cas no balancejat.

La classe positiva obté resultats lleugerament superiors en Recall i F1-score en comparació amb la classe negativa. Totalment diferent que en el cas amb resultats de dades no balancejades, on hi havia molt bias cap a la classe negativa.

Fold	Recall_0	Recall_1	Precision_0	Precision_1	F1_0	F1_1	Accuracy
1	0.29	0.98	0.93	0.57	0.45	0.72	0.6274
2	0.36	0.98	0.95	0.57	0.52	0.72	0.6442
3	0.68	0.66	0.67	0.67	0.67	0.67	0.6697
4	0.86	0.30	0.50	0.72	0.63	0.42	0.5506
5	0.31	0.95	0.86	0.58	0.46	0.72	0.6323
Mean ± std	0.5 ± 0.26	0.77 ± 0.29	0.78 ± 0.19	0.62 ± 0.06	0.55 ± 0.09	0.65 ± 0.13	0.62 ± 0.04

Figura 13. Resultats folds i mean ± std del model backbone en test NW balancejat

Aquest equilibri també es veu reflectit en la matriu de confusió acumulada on les prediccions per les dues classes estan molt més equilibrades, com es pot observar en la figura 14. Els TP i TN són similars, indicant que s'ha predit amb més justícia les dues classes. No obstant, fins i tot, en aquest cas està esbiaixat una mica per a la classe positiva.

41.527	43.256
20.360	64.423

Figura 14. Matriu Confusió Acumulada Backbone en test NW balancejat

En resum, amb dades balancejades, el model Backbone aconsegueix resultats més equilibrats entre les classes 0 i 1. Tot i que encara hi ha certa variabilitat entre folds, el balanç de dades ha millorat la capacitat del model per generalitzar de manera justa les dues classes. No obstant, els valors són més baixos.

4.3 Experiment 4

Les mètriques d'entrenament (figura 15) són prou bones i similars als experiments anteriors, tot i que es disminueix una mica la mitjana de l'Accuracy (0.7553) i el F1-score (0.7604). Mentre que l'AUC és lleugerament superior (0.8431). A més, la baixa variabilitat en la desviació estàndard mostra una bona consistència en els resultats en els diferents conjunts de validació.

	Accuracy	F1_score	AUC
Backbone	0.7553 ± 0.0532	0.7604 ± 0.0186	0.8431 ± 0.0285

Figura 15. Resultats mean ± std del model backbone en l'entrenament NI

En el test, el recall negatiu mitjà és de 0.83 amb una desviació estàndard de 0.08, el qual indica que el model és bastant bo a l'hora de detectar instàncies de la classe negativa, amb poca variabilitat entre els diferents folds, com es pot observar en la figura 16. Mentre que el recall per a la classe positiva és de 0.59, però amb una desviació estàndard superior, el que indica un rendiment més baix en la detecció dels atacs epilèptics. Aquests valors reflecteixen que el model està cometent falsos negatius.

Fold	Recall_0	Recall_1	Precision_0	Precision_1	F1_0	F1_1	Accuracy
1	0.84	0.45	0.77	0.57	0.81	0.50	0.72
2	0.72	0.66	0.71	0.66	0.71	0.66	0.69
3	0.88	0.71	0.87	0.74	0.88	0.72	0.83
4	0.77	0.56	0.78	0.55	0.78	0.55	0.7
5	0.94	0.56	0.85	0.77	0.89	0.65	0.83
Mean ± Std	0.83 ±	0.59 ±	0.80 ±	0.66 ±	0.81 ±	0.62 ±	0.75 ±
	0.08	0.10	0.07	0.09	0.07	0.08	0.07

Figura 16. Resultats folds i mean ± std del model backbone en test NI

Pel que respecta a la precisió, el 80% dels casos que prediu que no hi ha atac, ho fa correctament, però els casos positius són correctament detectats un 66% dels cops. El F1 score de la classe 0 és de 0.81 de mitjana, amb un bon equilibri entre la precisió i el recall i poca variabilitat entre els folds. En els casos positius tenim un F1-score mitjà de 0.62 reflectint un rendiment moderat en la classe dels atacs.

Observant aquestes mètriques i la matriu de confusió acumulada (figura 17), veiem que el model té un bon rendiment per a la classificació de la classe negativa, però el seu rendiment en la classe positiva podria millorar significativament. Tot i que té una bona capacitat per reduir els falsos positius i té un accuracy més elevat que l'experiment anterior, en una aplicació de detecció d'atacs epilèptics, és necessari millorar la capacitat del model per detectar els casos positius.

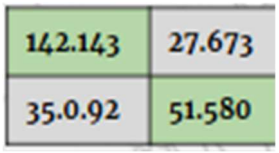


Figura 17. Matriu Confusió Acumulada Backbone en test NI

4.4 Experiment 5

Els resultats de test (figura 18) mostren que no hi ha tanta variabilitat en els folds per la classe positiva, a diferència dels folds de la classe negativa, ja que la desviació estàndard en la classe positiva és 0.10 de mitjana, en canvi, per la classe negativa és de 0.25. Aleshores hi ha molta variabilitat en els folds dins de la classe negativa. Cosa que també es pot observar, per exemple, en la precisió del fold 1 que és 1.00, però la del fold 2 és 0.52.

A més, continua tenint molt desequilibri entre classes com en els altres experiments. En aquest experiment, el model prioritza la classe positiva, possiblement perquè és la classe més freqüent. Aquesta classe té mètriques més altes especialment en Recall (0.89) i F1-score(0.73) en comparació a la classe negativa amb 0.43 i 0.52, la qual cosa implica que és com a l'atzar la predicció.

Fold	Recall_0	Recall_1	Precision_0	Precision_1	F1_0	F1_1	Accuracy
1	0.54	1.00	1.00	0.68	0.70	0.81	0.7681
2	0.11	0.90	0.52	0.50	0.18	0.64	0.5041
3	0.59	0.74	0.70	0.65	0.64	0.69	0.6691
4	0.72	0.94	0.92	0.77	0.81	0.85	0.8297
5	0.17	0.86	0.55	0.51	0.26	0.64	0.5155
Mean ± std	0.43 ± 0.27	0.89 ± 0.10	0.74 ± 0.22	0.62 ± 0.12	0.52 ± 0.28	0.73 ± 0.10	0.66 ± 0.15

Figura 18. Resultats folds i mean ± std del model backbone en test NP

Aquest comportament també s'observa en la matriu de confusió acumulada. On es veu que el model està esbiaixat cap a la classe 1, com es pot observar en la figura 19.

5.956	9.532
1.635	13.853

Figura 19. Matriu Confusió Acumulada Backbone en test NP

5. Conclusions i millores

Podem concloure que hem dut a terme les tasques del projecte amb dedicació, implementant dues arquitectures per a la detecció de patrons: flatten i backbone amb LSTM. A més, hem realitzat l'entrenament i la validació de tres dissenys experimentals per analitzar el rendiment del nostre model.

Tanmateix, els resultats obtinguts no són del tot satisfactoris, ja que presenten valors desequilibrats i baixos. Entre les possibles millores proposades, considerem important revisar el balanceig de les dades en les proporcions dels k-folds. També caldria tenir en compte la mida de les seqüències, aspecte que hem passat per alt en centrar-nos només en el batch_size. A més, seria recomanable incrementar el nombre d'èpoques, ja que en aquest projecte ens hem limitat a quaranta. On un nombre més elevat d'èpoques podria haver millorat el rendiment del model i incrementat les mètriques d'avaluació. Aquestes podrien ser algunes de les raons principals dels nostres resultats desequilibrats i poc satisfactoris.

Finalment, cal destacar que les dificultats associades als problemes de memòria han condicionat significativament el desenvolupament del projecte.

6. Referencies

[1] Hernández-Sabaté A, Yauri J, Folch P, Piera MÀ, Gil D. Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. Applied Sciences.

[2] S. Chakrabarti, A. Swetapadma, P. K. Pattnaik, A Channel Independent Generalized SeizureDetection Method for Pediatric Epileptic Seizures, Computer Methods and Programs in Biomedicine

[3] M. Shroff. "Know your Neural Network architecture more by understanding these terms". Medium. [En línea]. Disponible: https://medium.com/@shroff-megha6695/know-your-neural-network-architecture-more-by-understanding-these-terms-67faf4ea0efb

[4] B. Nalawade. "The Essential Guide to K-Fold Cross-Validation in Machine Learning". Medium. Accedido el 21 de enero de 2025. [En línea]. Disponible: https://medium.com/@bididudy/the-essential-guide-to-k-fold-cross-validation-in-machine-learning-2bcb58c50578