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

Viera Antequera, Jordi. *Predicció del risc de dislèxia en infants de 3r d'educació primària mitjançant l'anàlisi de la lectura oral amb IA*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334384>

under the terms of the  license.

Predicció del risc de dislèxia en infants de 3r d'educació primària mitjançant l'anàlisi de la lectura oral amb IA

Jordi Viera Antequera

Resum—Aquest treball investiga l'aplicació de tècniques d'intel·ligència artificial per detectar el risc de dislèxia en infants de 3r de primària a partir d'enregistraments de veu llegint un text. S'utilitzen paràmetres generats amb *faster-whisper* per capturar errors de lectura i fluïdesa, que es complementen amb descriptors acústics profunds del model HuBERT. Per afrontar el fort desequilibri de la mostra real, s'avalua l'ús de penalitzacions de pes en els algorismes. Així mateix, s'experimenta amb una cerca combinada de paràmetres i selecció de variables per a models SVM, Random Forest i XGBoost, ajustant els llindars de decisió final per maximitzar la capacitat de detecció preventiva del sistema. Els resultats obtinguts demostren que l'enfocament multimodal optimitzat mitiga l'efecte del desequilibri de dades i assoleix una elevada sensibilitat en la detecció, establint una base sòlida per al desenvolupament futur d'eines automatitzades de suport al cribratge escolar.

Paraules clau—Cribratge escolar, dislèxia, HuBERT, *Machine Learning*, processament d'àudio, Whisper.

Abstract—This study investigates the application of artificial intelligence techniques to detect dyslexia risk in 3rd-grade children using voice recordings of them reading a text. Features generated with *faster-whisper* are utilized to capture reading and fluency errors, which are complemented by deep acoustic descriptors extracted with the HuBERT model. To address the severe imbalance of the real sample, the use of class-weight penalties within the algorithms is evaluated. Likewise, the study experiments with a combined parameter search and variable selection for SVM, Random Forest, and XGBoost models, specifically adjusting their final decision thresholds to maximize the system's preventive detection capabilities. The results demonstrate that the optimized multimodal approach mitigates the effect of data imbalance and achieves high detection sensitivity, establishing a solid foundation for the future development of automated tools to support school screening.

Index Terms—Audio processing, dyslexia, HuBERT, *Machine Learning*, school screening, Whisper.

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

ELS trastorns del neurodesenvolupament, com la dislèxia, afecten significativament l'adquisició de les habilitats lectores durant l'educació primària [1]. Aquest trastorn específic de l'aprenentatge es manifesta en dificultats per reconèixer paraules de manera precisa i fluida, així com en capacitats deficientes de descodificació i lletreig, malgrat presentar una intel·ligència dins de la normalitat o superior. En els infants de 3r de primària, aquesta etapa és crítica per a la detecció precoç: l'alumne ja ha estat exposat a l'ensenyament formal de la lectura, però encara no ha desenvolupat estratègies de compensació que puguin ocultar el trastorn, fet que facilita una intervenció psicopedagògica eficaç.

Actualment, la identificació de la dislèxia es basa en

observacions a l'aula i en bateries de testos manuals. No obstant això, aquests mètodes tradicionals són laboriosos, lents i introdueixen subjectivitat. En aquest context, la intel·ligència artificial i el processament de la parla ofereixen eines de suport per analitzar àudios de manera automàtica [2], aportant indicadors objectius, no invasius i escalables.

Tot i els avenços, la majoria de la recerca s'ha centrat en llengües com l'anglès o el castellà. A més, el salt cap a entorns escolars reals presenta reptes metodològics importants, principalment derivats del fort desequilibri entre els alumnes sans i els casos de risc en mostres naturals.

Aquest treball planteja l'ús d'aprenentatge automàtic per analitzar àudios de lectura d'infants de 3r de primària de parla catalana. L'objectiu és explorar el potencial combinat de descriptors textuais i característiques acústiques profundes extretes de la veu per predir el risc de dislèxia. A diferència d'enfocaments basats en dades artificials, aquesta recerca aposta per un tractament net de la informació real recollida a les escoles, adaptat a les

- E-mail de contacte: jviera.antequera@gmail.com
- Menció realitzada: Computació
- Treball tutoritzat per: Marta Prim Sabrià (Dep. Microelectrònica i Sistemes Electrònics)
- Curs 2025/26

necessitats de sensibilitat del cribratge educatiu.

El document s'estructura partint d'una revisió dels objectius i del marc tecnològic actual. A continuació, es descriuen el conjunt de dades escolar i la metodologia de preprocessament i extracció multimodal de característiques. Finalment, es detallen el protocol experimental, l'anàlisi de resultats en els diferents escenaris i les conclusions i línies de futur.

2 OBJECTIUS

L'objectiu d'aquest projecte és identificar indicadors de risc de dislèxia en infants a partir d'enregistraments d'àudio llegint un text normalitzat en llengua catalana. Per aconseguir-ho, s'aplicaran algorismes d'aprenentatge automàtic per desenvolupar un model capaç d'identificar la probabilitat que un alumne presenti aquest trastorn. El propòsit final és crear una eina de suport objectiva, no invasiva i escalable per als professionals de l'àmbit educatiu i de l'orientació —com docents, logopedes o psicopedagogs— que faciliti un cribratge precoç a les aules.

3 ESTAT DE L'ART

Davant les limitacions de les avaluacions clíniques tradicionals —que consumeixen molt de temps, són costoses i depenen de personal altament especialitzat—, la recerca en intel·ligència artificial ha explorat diferents vies tecnològiques per automatitzar el cribratge de la dislèxia [5]. Aquests enfocaments es classifiquen en tres grans categories segons la naturalesa de les dades recollides.

3.1 Sistemes de seguiment ocular (*Eye-tracking*)

Aquesta línia analitza els moviments dels ulls de l'infant durant la lectura. Mitjançant aprenentatge automàtic, es processen patrons espaiotemporals com les fixacions (temps que la mirada es deté), les sacades (saltos entre paraules) i les regressions (retorns a fragments ja llegits). Tot i que mesuren directament la càrrega cognitiva, requereixen un maquinari de captació específic, calibratges complexos i un entorn controlat, fet que en dificulta el desplegament massiu a les escoles.

3.2 Sensors neurofisiològics

Aquest enfocament estudia l'activitat biològica del sistema nerviós central durant la lectura. Utilitza dades d'electroencefalografies (EEG) per mesurar canvis elèctrics cerebrals en temps real, o ressonàncies magnètiques (fMRI/MRI) per detectar activitat anòmala en les àrees del processament fonològic [5]. Malgrat la seva profunditat clínica, són mètodes invasius i d'elevat cost econòmic i logístic, inviablès per a campanyes preventives a gran escala.

3.3 Anàlisi d'àudio de lectura

Com a alternativa, l'anàlisi de senyals de parla s'ha consolidat com la via més prometedora per a l'entorn escolar [3]. Es basa en l'enregistrament digital de la veu dels infants llegint llistes de paraules o textos per analitzar automàticament variables com els errors de pronunciació, la durada de les pauses, la velocitat i el ritme prosòdic [4]. És una aproximació no invasiva, de baix cost i executable amb micròfons convencionals ja presents en els dispositius escolars, establint-se com la solució òbvia per a un cribratge massiu.

És precisament en aquesta línia on s'enfoca el present treball, utilitzant la gravació de veu com a l'únic canal d'entrada per a l'anàlisi.

4 METODOLOGIA I PLANIFICACIÓ

El flux de treball s'estructura en un procés incremental dissenyat per transformar els enregistraments originals en un model de detecció precoç mitjançant les següents fases executives ordenades:

En primer lloc, es realitza una avaluació inicial del conjunt d'àudios per descartar aquells enregistraments tallats o amb deficiències crítiques de gravació que n'impedeixen la comprensió. Superat aquest filtre, el **processament de senyal i extracció multimodal** s'encarrega de netejar i homogeneïtzar els àudios per extreure les característiques textuais, de fluïdesa lectora i les representacions acústiques profundes de la parla. A continuació, l'**anàlisi exploratòria de dades** permet comprendre la distribució de les variables i la seva relació estadística amb les dificultats lectores. Seguidament, la **fusió i integració de dades** combina de manera híbrida les diferents fonts d'informació (text i so), optimitzant l'espai de característiques per evitar redundàncies.

Un cop preparat el conjunt de dades, es procedeix al **modelatge i optimització d'hiperparàmetres**, on s'entrenen diversos algorismes aplicant una cerca en graella (*GridSearchCV*) amb validació creuada per avaluar sistemàticament les millors combinacions de paràmetres. Finalment, l'**optimització del llindar de cribratge** ajusta el sistema de decisió final per prioritzar la capacitat del model per detectar correctament els infants en risc i minimitzar els errors d'omissió (falsos negatius). Per a la realització de tot aquest procés, la planificació temporal i el cronograma seguit es detallen a l'Apèndix 1.

5 EL CONJUNT DE DADES (*DATASET*)

5.1 Descripció de la mostra escolar

Els enregistraments d'àudio utilitzats en aquest projecte han estat proporcionats per la tutora del treball, obtinguts en col·laboració amb centres educatius de Catalunya sota consentiment formal dels tutors legals. Per tractar-se de

dades altament sensibles de menors d'edat, la informació és completament anònima i d'ús exclusivament acadèmic, complint amb les regulacions vigents de protecció de dades.

La mostra final està formada per 143 enregistraments d'àudio independents, recollits en quatre centres escolars diferents, corresponents a la lectura en veu alta d'un text breu i normalitzat en llengua catalana titulat "En Carles" (vegeu l'Apèndix 2). Totes les mostres provenen d'alumnes de 3r de primària (entorn dels 8 anys d'edat), en què la mecànica lectora bàsica ja hauria d'estar consolidada. L'etiqueta de referència (*ground truth*) de cada participant està codificada en la nomenclatura del mateix fitxer, determinant si l'infant compta amb un diagnòstic o indicatiu clar de risc de dislèxia confirmat prèviament per especialistes.

5.2 Desequilibri de classes

En analitzar la distribució real del dataset mitjançant l'etiqueta binària, s'evidencia una forta asimetria que reflecteix fidelment la proporció d'aquest trastorn a les aules. Del total de 143 alumnes, 128 (89,5%) pertanyen al grup de control o sense risc, mentre que només 15 infants (10,5%) estan identificats dins del grup de risc de dislèxia, tal com es detalla a la Figura 1.

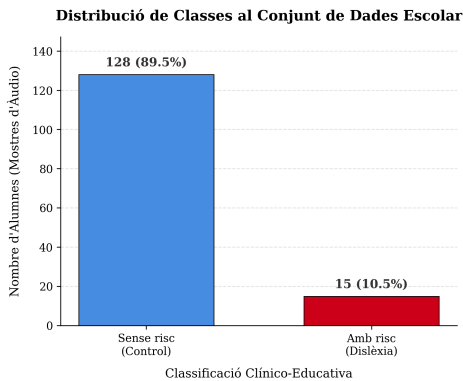


Fig. 1. Distribució d'alumnes segons el risc de dislèxia.

Aquest desequilibri de classes (un ràtio aproximat de 9 a 1) imposa un repte metodològic central. Els classificadors d'aprenentatge automàtic convencionals tendeixen a optimitzar l'exactitud global (*accuracy*) col·lapsant cap a la classe majoritària (control), fet que generaria un volum inacceptable de falsos negatius en un entorn escolar real.

6 PROGRAMARI I LLIBRERIES

El desenvolupament s'ha centralitzat en Visual Studio Code utilitzant el llenguatge Python. Per optimitzar l'espai i garantir un flux de treball eficient, les llibreries utilitzades s'organitzen segons els següents blocs funcionals:

- **Tractament de senyal i àudio:** S'ha emprat *pydub*

per a la manipulació de fitxers i detecció de silencis, i *librosa* per a la càrrega, anàlisi i extracció de propietats acústiques de la veu.

- **Deep Learning i extracció de característiques:** S'utilitza *faster-whisper* per a la transcripció automatitzada d'alta eficiència, i *transformers* (Hugging Face) per carregar el model preentrenat HuBERT i generar els *embeddings* acústics profunds.
- **Mètriques de text:** *difflib* [10] s'encarrega de l'alineació de text i comparació de cadenes, mentre que *jiwer* calcula de manera automatitzada les taxes d'error de paraula (WER) i de caràcters (CER) respecte al text patró.
- **Ciència de dades i classificació:** *pandas* i *numpy* s'utilitzen per a la transformació de dades tabulars, i *scikit-learn* per a la normalització (StandardScaler), reducció de dimensionalitat i selecció de característiques (PCA i RFE). La validació creuada s'executa amb *StratifiedKfold* per preservar l'asimetria de les classes. Els algorismes de modelatge utilitzats són *RandomForestClassifier*, *SVC* i *XGBClassifier*, avaluant el seu rendiment mitjançant el mòdul *metrics* (*Accuracy*, *Precision*, *Recall*, *F2-Score* i *AUC*).
- **Visualització de dades:** *matplotlib* i *seaborn* s'encarreguen de la generació de gràfiques de qualitat acadèmica, com les corbes ROC i les matrius de confusió.

7 PREPROCESSAMENT I EXTRACCIÓ DE CARACTERÍSTIQUES

Abans d'introduir els enregistraments d'àudio als models d'intel·ligència artificial, la informació bruta s'ha de netejar, estandarditzar i transformar en variables numèriques homogènies. Aquest procés s'executa en tres fases seqüencials: una fase inicial de neteja i normalització de l'àudio, una fase paral·lela d'extracció (textual amb *faster-whisper* i acústica profunda amb HuBERT), i una fase final d'integració de dades, tal com es mostra a la Figura 2.

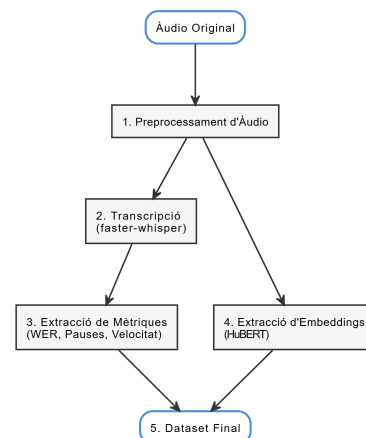


Fig. 2. Arquitectura del sistema.

7.1 Estandardització de l'àudio

Els àudios originals mostraven variabilitat en formats i condicions acústiques segons l'entorn escolar. Per homogeneïtzar-los, es va dissenyar un procés automatitzat amb *pydub* on cada pista es converteix a canal mono i es remostreja a 16.000 Hz, estàndard natiu dels models posteriors. A més, s'aplica una normalització d'amplitud a -20,0 dBFS per uniformitzar el volum. Finalment, un sistema de Detecció d'Activitat de Veu (VAD) elimina el soroll de fons i silencis per sota de -35 dBFS, retallant l'àudio després de l'última paraula.

7.2 Transcripció automàtica

Un cop homogenis els fitxers d'àudio, s'utilitza l'arquitectura *faster-whisper* carregant el model *Whisper-large-v3-ca-punctuated-3370h* del projecte AINA [6]. Aquest model genera de manera nativa transcripcions en català amb puntuació i majúscules.

Durant aquesta fase, es va detectar la presència d'**al·lucinacions** al final de molts enregistraments (bucles de paraules o frases inventades pel model en trobar silencis o soroll de fons) [8], [9]. Per solucionar-ho i evitar distorsions en el càlcul d'errors, es va dissenyar un sistema de truncat automàtic mitjançant *SequenceMatcher* de la llibreria *difflib* [10], estructurat en tres passos:

1. **Identificació de blocs coincidents:** Es comparen les paraules de la transcripció bruta amb les del text original de referència.
2. **Cerca del punt de tall:** L'algorisme localitza l'última posició vàlida on hi ha hagut una coincidència real, marcant el final de la lectura de l'infant.
3. **Eliminació de text sobrant:** Es descarta de forma directa tot el text posterior a aquesta posició, eliminant per complet les al·lucinacions.

D'altra banda, per solucionar la manca d'homogeneïtat dels alumnes que havien llegit el títol del text i els que no, s'aplica un filtre automàtic que detecta la primera paraula de la història i elimina el títol si s'ha llegit. Finalment, tot i que la puntuació de *Whisper* és clau per a les pauses, distorsionava la comparació de paraules. Per resoldre aquest biaix i assegurar una comparació justa, el text es normalitza mitjançant la conversió completa a minúscules i l'eliminació absoluta de qualsevol signe de puntuació, garantint que les futures mètriques es basin només en desviacions reals de la lectura.

7.3 Extracció de mètriques lectores i temporals

A partir de la transcripció normalitzada, els marcadors de *Whisper* i l'anàlisi de silencis de *pydub*, es construeix un vector tabular d'indicadors dividit en tres dimensions de diagnòstic:

- **Mètriques de precisió i error:** Avaluen la fidelitat lectora mitjançant el WER (*Word Error Rate*) i el

CER (*Character Error Rate*), que mesuren la taxa de paraules i caràcters afectats per omissions o substitucions. S'afegeix la variable *Error_Ponderat*, que assigna un pes superior als errors en paraules de major complexitat o longitud (vegeu l'Apèndix 2).

- **Mètriques de temporalitat i pauses:** Quantifiquen el cost cognitiu a través de la segmentació del senyal. Es compon de *Temps_sense_silencis* (durada neta de la veu), *Proporció_lectura_silenci* (relació percentual entre veu i silenci), i el recompte i mesura de vacil·lacions (*Nombre_pauses*, *Temps_total_pausa* i *Durada_mitjana_pausa*).
- **Mètriques de velocitat i ritme:** Mesuren el grau d'automatització de la descodificació fonològica de l'alumne mitjançant la *Duració_sec* (temps total de l'exercici), *Velocitat_lectora_pal_sec* (paraules per segon) i *Taxa_paraules* (ritme net de generació verbal).

D'aquesta manera, l'espai de característiques textuais queda definit per un total d'11 indicadors quantitatius.

7.4 Extracció d'embeddings

Més enllà de la informació textual, el senyal acústic aporta descriptors clau sobre el comportament lector. Per capturar-los, s'utilitza el model *HuBERT Base (BSC-LT/hubert-base-ca-2k)*, preentrenat específicament en català en el marc del projecte AINA [6]. Aquest model transforma el senyal d'àudio en vectors numèrics que encapsulen característiques prosòdiques, tonals i de ritme a través de tres etapes:

- **Extracció de vectors (Mean Pooling):** El pas de l'àudio pel model genera una matriu de 768 dimensions per cada fragment temporal (*frame*). Per reduir la redundància temporal i obtenir una representació homogènia, s'aplica una tècnica de *mean pooling* que fa la mitjana de la seqüència, consolidant la informació en un únic vector final de 768 dimensions per a cada àudio [7].
- **Reducció dimensional amb PCA:** Amb l'objectiu d'evitar el sobreajust (*overfitting*) provocat per l'alta dimensionalitat (768 variables), s'aplica una Anàlisi de Components Principals (PCA) [11]. En lloc de fixar un nombre de components arbitrari, s'implementa un criteri de selecció dinàmic que determina automàticament el nombre mínim de components necessaris per retenir el 85% de la variància original del senyal acústic.
- **Anàlisi de convergència:** Aquest procés es valida mitjançant la corba de variància acumulada (vegeu la Figura 3), que permet verificar de manera visual el punt de tall on la informació esdevé pràcticament constant, mantenint finalment un total de 13 components principals. Aquesta reducció comprimeix l'espai acústic eliminant el soroll estadístic i facilitant la

convergència dels classificadors posteriors.

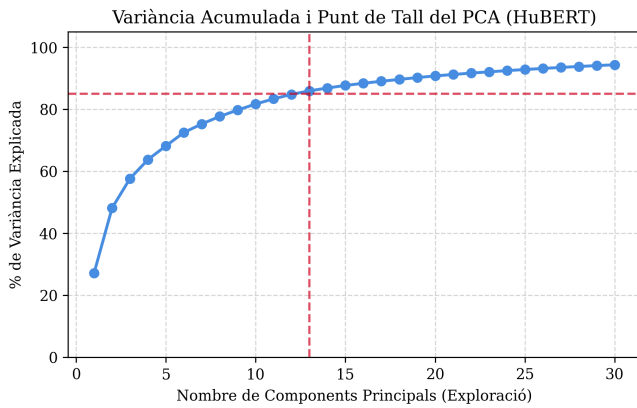


Fig. 3. Variància acumulada del PCA sobre els embeddings de HuBERT i selecció del punt de tall (85%).

8 DISSENY DE L'EXPERIMENT I MODELATGE

Una vegada consolidat el conjunt de dades, s'ha dissenyat un protocol experimental per entrenar i validar els models de classificació. Aquesta fase s'ha estructurat com una estratègia incremental de millora, orientada a avaluar el comportament basal de diferents fonts d'informació i a optimitzar progressivament els algorismes per predir el risc de dislèxia en un entorn d'alta asimetria.

8.1 Algorismes de classificació considerats

Amb l'objectiu de trobar el sistema de predicció més robust, es comparen tres algorismes d'aprenentatge automàtic reconeguts per la seva eficàcia en dades tabulars: *Random Forest*, Màquines de Vectors de Suport (*Support Vector Machine, SVM*) i *XGBoost*.

S'ha optat per aquests models i s'han descartat algorismes més tradicionals com la Regressió Logística o els Arbres de Decisió simples per dues raons principals:

- **Capacitat de capturar relacions complexes:** La dislèxia es manifesta a través d'una combinació subtil d'errors de text i variacions acústiques. Els models lineals assumeixen independència i relacions lineals, fet que impedeix detectar patrons creuats complexos (per exemple, que un text amb poques pauses però amb un to de veu molt variable indiqui risc).
- **Robustesa davant el soroll:** Els models basats en conjunts d'arbres i les SVM amb kernels no lineals gestionen millor les dades sorolloses dels àudios escolars reals en comparació amb les alternatives clàssiques.

Per avaluar l'eficàcia real de cada font d'informació de

manera aïllada i combinada, els tres models s'executen i es comparen sota quatre escenaris experimentals diferents:

- **Configuració A (Textual):** Utilitza exclusivament les mètriques derivades de la transcripció i l'alineació de text de *faster-whisper*.
- **Configuració B (Acústica):** S'entrena únicament amb les dades de so codificades en els components de HuBERT.
- **Configuració C (Multimodal Completa):** Fusiona el conjunt complet de variables inicials (24 descriptors).
- **Configuració D (Optimitzada):** Empra un subconjunt seleccionat i optimitzat mitjançant tècniques avançades de reducció i cerca d'hiperparàmetres.

8.2 Criteri d'avaluació

Per seleccionar els millors escenaris i sintonitzar els paràmetres de cada algorisme, és imprescindible definir una mètrica objectiu que s'ajusti a la naturalesa asimètrica de les dades i, sobretot, a la finalitat pràctica de la recerca. En tasques de classificació estàndard, s'utilitza habitualment la mètrica *F1-Score*, la qual calcula la mitjana harmònica ponderada equitativament entre la *Precision* (precisió) i el *Recall* (sensibilitat). No obstant això, atès que aquest sistema s'ha concebut com una eina de cribratge preventiu en l'entorn escolar, els costos d'error són altament asimètrics:

- **Cost d'un fals negatiu (més crític):** Classificar un infant com a control quan realment té risc de dislèxia implica que quedarà fora del radar educatiu i no rebrà la intervenció precoç necessària.
- **Cost d'un fals positiu (menys crític):** Identificar preventivament un infant com a "cas de risc" simplement suposa que passarà a una segona fase d'avaluació individualitzada per l'equip psicopedagògic, el qual podrà descartar el cas sense conseqüències greus.

Amb l'objectiu de traslladar aquesta prioritat a la lògica de l'aprenentatge automàtic, es descarta l'ús de l'*F1-Score* tradicional i s'adopta l'**F2-Score**. Aquest coeficient es formula matemàticament de la següent manera:

$$F_2 = \frac{5}{\frac{4}{Precision} + \frac{1}{Recall}} = \frac{5}{\frac{4TP+4FN}{TP} + \frac{TP+FP}{TP}}$$

Mitjançant aquest coeficient, la funció d'optimització

atorga el doble de pes al Recall que a la Precision. D'aquesta manera, tots els experiments posteriors (tant el tractament del desequilibri com el procés de selecció de variables i la cerca en graella) estaran orientats de manera algorísmica a protegir la detecció de la classe minoritària, evitant l'omissió d'infants amb dificultats lectores.

8.3 Tractament del desequilibri

Com que el conjunt de dades presenta un fort desequilibri de classes (on els casos de risc representen només el 10,5% del total), s'ha implementat una estratègia de penalització matemàtica directa en els algorismes de classificació per evitar el biaix cap a la classe majoritària durant les proves inicials de les Configuracions A, B i C.

Aquesta tècnica consisteix a calcular la raó de desequilibri real i integrar-la en la funció de cost dels models. D'aquesta manera, cometre un error de predicció amb un infant de la classe de risc (dislèxia) penalitza significativament més que cometre'l amb un de la classe de control, forçant els algorismes a prioritzar els patrons característics de la dislèxia de forma estrictament algorísmica.

L'impacte d'aquesta penalització s'ha avaluat de manera pràctica en els tres models (Figura 4). Els resultats mostren que tant la SVM com el Random Forest milloren de manera radical quan s'activen els pesos (balanced). Per contra, l'algorisme XGBoost (configurat mitjançant `scale_pos_weight`) mostra un canvi molt més moderat. Això passa perquè l'XGBoost aprèn en sèrie i cadascun dels seus arbres es concentra a corregir els errors de l'anterior; aquesta naturalesa el fa resiliència al desequilibri per si mateix, de manera que l'ajust de pesos el beneficia, però sense causar un salt tan dràstic.

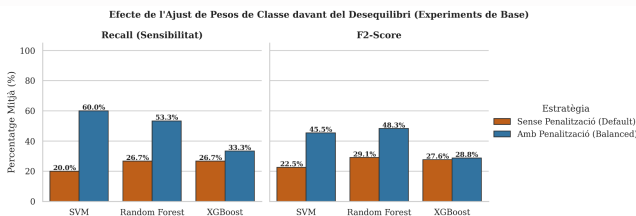


Fig 4. Impacte de la penalització de classe en SVM, Random Forest i XGBoost. Es compara el rendiment mitjà (%) en Recall i F2-Score

Per garantir la robustesa metodològica, tot el procés d'avaluació s'executa mitjançant una validació creuada estratificada en 5 particions (*Stratified 5-Fold Cross-Validation*). Aquest procediment assegura que la proporció real de la mostra es mantingui constant tant en els conjunts d'entrenament com en els de test de cada iteració, evitant estimacions optimistes o esbiaixades del rendiment global.

8.4 Selecció i reducció de característiques

Per a l'escenari final (Configuració D), el vector inicial comptava amb 24 variables numèriques. Tot i la reducció prèvia amb PCA sobre els embeddings acústics, moltes característiques combinades presentaven una alta col·linearitat, fet que introduïa redundància i soroll estadístic amb risc de sobreajust (*overfitting*). Per resoldre-ho, es va dissenyar una optimització espacial en tres fases.

En la primera fase, comuna per a tots els enfocaments, es va calcular la matriu de correlació de Pearson (vegeu l'Apèndix 3) amb un llindar crític del 85%. Aquest filtre va revelar fortes dependències en el bloc temporal i de ritme, eliminant automàticament cinc variables: *Velocitat_lectora_pal_sec*, *CER*, *Temps_sense_silencis*, *Temps_total_pausa* i *Durada_mitjana_pausa*.

Amb les 19 variables restants, s'aplica l'algorisme d'Eliminació Recursiva de Característiques (RFE) per extreure la combinació òptima adaptada a cada arquitectura. Aquest mètode de selecció inversa (*backward selection*) avalua el rendiment de forma iterativa, eliminant en cada ronda la variable amb menys pes intern fins a trobar el subconjunt òptim. L'aplicació de l'RFE es va diferenciar en dues rutes segons la naturalesa matemàtica de cada model per garantir el rigor experimental:

- **Ruta per a models d'arbres (Random Forest i XGBoost):** L'RFE es va executar utilitzant la importància de les característiques (*feature importances*) intrínseca dels arbres. Aquest procés va permetre identificar un subconjunt estàtic i homogeni de **7 variables optimitzades** que maximitzaven l'F2-score en aquestes estructures.

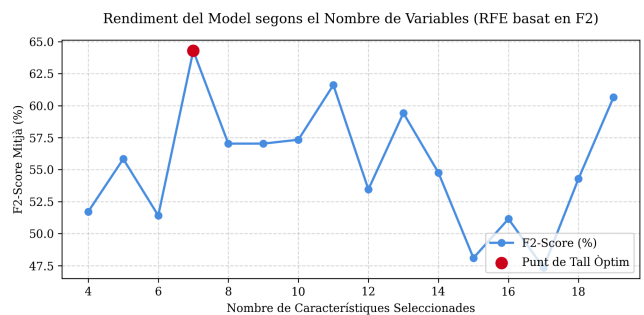


Fig. 5. Evolució de l'F2-score mitjà en funció del nombre de variables seleccionades mitjançant RFE, destacant el punt de tall òptim.

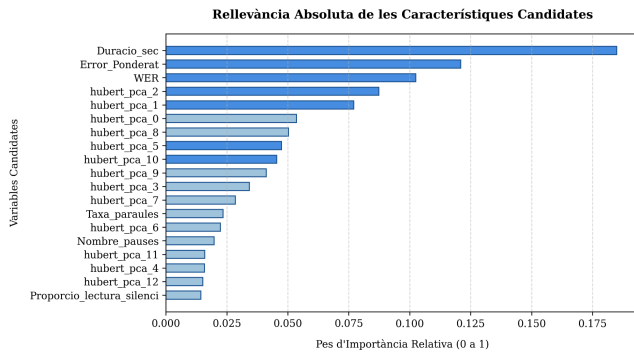


Fig. 6. Importància relativa de les variables seleccionades després de l'optimització.

- **Ruta per a Màquines de Vectors de Suport (SVM):** Per al model SVM, l'algorisme RFE s'ha integrat estrictament a l'interior de cada *fold* de la validació creuada. Com que la SVM és altament sensible a la distribució geomètrica, un RFE global hagués escollit variables coneixent tota la mostra, inflant artificialment els resultats (*overfitting*). Al fer-ho dins de cada *fold*, les variables se seleccionen usant només el subconjunt d'entrenament, cosa que garanteix que les dades de test actuïn com una mostra completament nova i desconeguda per al model.

Aquesta cerca iterativa interna va utilitzar els pesos de l'SVM Lineal per avaluar la importància de les variables de forma directa. Els subconjunts seleccionats a cada plec es van fer servir posteriorment per entrenar també el model amb kernel de Base Radial (RBF). Tot i que l'RBF treballa en un espai no lineal on no es poden extreure coeficients directes, les variables rellevants en l'entorn lineal solen mantenir el seu poder discriminant en ser transformades pel kernel, aconseguint així un model robust sense multiplicar el cost de càlcul.

Un cop definit el comportament de reducció espacial de la Configuració D, s'aplica una cerca en graella (GridSearchCV) acoblada a la validació creuada per avaluar de manera sistemàtica múltiples combinacions de paràmetres, garantint que cada model s'ajusti a la seva màxima capacitat operativa sota l'objectiu d'optimitzar l'*F2-Score*. L'espai explorat per a cada algorisme es detalla a la **Taula 1**:

TAULA 1
HIPERPARÀMETRES PROVATS EN ELS MODELS

Paràmetre	Descripció	Valors provats
Random Forest		
n_estimators	Nombre d'arbres	[50, 100, 200]
max_depth	Profunditat màxima	[3, 5, None]
min_samples_split	Mostres mínimes per dividir un node	[2, 5]
XGBoost		
n_estimators	Nombre d'arbres	[100, 150, 200]
learning_rate	Taxa d'aprenentatge	[0.01, 0.03, 0.1]

max_depth	Profunditat màxima	[3, 4, 5]
subsample	Fracció de mostres per entrenar cada arbre	[0.7, 0.8, 1.0]
reg_alpha	Terme de regularització L1	[0.0, 0.1, 0.5]
reg_lambda	Terme de regularització L2	[1.0]
SVM		
kernel	Tipus de funció de mapa geomètric	['linear', 'rbf']
C	Paràmetre de penalització de l'error	[0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1]
n_features	Nombre de variables seleccionades	[5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19]

8.5 Ajust de l·lindars per a cribratge preventiu

Com a darrer pas del disseny experimental, i atès que l'objectiu final de l'eina és el cribratge preventiu en l'entorn escolar, s'aplica una estratègia de decisió final adaptada a la naturalesa matemàtica de cada model per tal de prioritzar el *Recall* i evitar deixar infants de risc sense identificar:

- **Algorismes basats en arbres (Random Forest i XGBoost):** Es modifica el l·lindar típic del 50% i es fixa en el **30%**. Random Forest genera la probabilitat final segons la proporció de vots positius dels arbres, i XGBoost mitjançant la suma ponderada de les decisions de cada iteració. A causa de l'alta asimetria de les dades, els models mostren una tendència inherent a beneficiar la classe majoritària, fet que dona lloc a prediccions de risc massa prudents. Fixar el tall al 30% compensa aquest biaix, permetent capturar els infants en risc al menor índex i prioritzant el cribratge.
- **Màquines de Vectors de Suport (SVM):** Es manté el l·lindar estàndard del **50%**. A diferència dels arbres, l'SVM no genera probabilitats per recomptes, sinó que construeix un hiperplà d'alta dimensió per separar geomàtricament les dues classes. Amb l'opció *balanced* activa, l'algorisme ja ajusta i desplaça internament aquest hiperplà durant l'entrenament per allunyar-lo de la classe majoritària. Com que la separació geomètrica ja està calibrada en la funció de cost, forçar un moviment posterior del l·lindar alteraria artificialment els marges, desestabilitzant el model i incrementant els falsos positius.

9 RESULTATS I DISCUSSIÓ

En aquest apartat es presenten i s'analitzen els resultats obtinguts. L'anàlisi s'estructura de manera incremental, partint del comportament basal de les fonts d'informació

aïllades i combinades (Configuracions A, B i C) fins a arribar a l'avaluació del model multimodal optimitzat (Configuració D).

9.1 Rendiment de les configuracions de referència (A, B i C)

Per establir la línia de partida de l'anàlisi, se seleccionen els resultats obtinguts en el millor cas de la validació creuada per a cadascun dels tres algorismes configurats amb els seus paràmetres per defecte i penalització per desequilibri de classe. L'avaluació independent i combinada d'aquestes fonts d'informació inicials revela comportaments molt diferenciats segons l'arquitectura de cada model.

En primer lloc, s'avalua la **Configuració A**, centrada exclusivament en les mètriques textuais i de fluïdesa extretes a partir de la transcripció de *faster-whisper*:

TAULA 2
RESULTATS DE LA CONFIGURACIÓ A (NOMÉS TEXTUAL)

Configuració A				
Model	Acc.	Prec.	Recall	F2-Score
Random Forest	0.888	0.420	0.533	0.501
XGBoost	0.825	0.255	0.467	0.393
SVM	0.895	0.000	0.000	0.000

En aquest escenari textual, el model Random Forest ofereix la millor capacitat de detecció primerenca, demostrant que els indicadors de ritme, taxes d'error i pauses temporals tenen un valor predictor immediat. Per contra, l'SVM col·lapsa per complet davant l'asimetria de la mostra, sent incapaç d'identificar cap cas de risc.

En segon lloc, la **Configuració B** analitza el rendiment utilitzant de forma aïllada els descriptors acústics profunds extrets mitjançant el model *HuBERT*:

TAULA 3
RESULTATS DE LA CONFIGURACIÓ B (NOMÉS ACÚSTICA)

Configuració B				
Model	Acc.	Prec.	Recall	F2-Score
Random Forest	0.916	0.650	0.400	0.425
XGBoost	0.866	0.450	0.533	0.510
SVM	0.888	0.067	0.533	0.067

A diferència del bloc anterior, la informació purament sonora interactua millor amb el model XGBoost, on s'assoleix l'*F2-Score* de partida més alt d'aquest escenari. Per contra, el Random Forest empitjora lleument el seu rendiment general respecte a les mètriques textuais. L'SVM deixa de presentar una incapacitat total per detectar la classe minoritària, però mostra un rendiment residual.

Finalment, la **Configuració C** posa a prova la fusió multimodal directa sumant la totalitat de les 24 variables d'entrada sense cap tipus de filtratge estructural:

TAULA 4
RESULTATS DE LA CONFIGURACIÓ C (MULTIMODAL COMPLETA)

Configuració C				
Model	Acc.	Prec.	Recall	F2-Score
Random Forest	0.902	0.547	0.547	0.533
XGBoost	0.846	0.280	0.280	0.467
SVM	0.874	0.000	0.000	0.000

Els resultats d'aquesta fusió simple són ambigus. Per al Random Forest s'observa una millora en l'estabilitat general del sistema, ja que manté la seva sensibilitat i eleva la precisió combinant text i so. No obstant això, l'XGBoost pateix un deteriorament generalitzat respecte al seu escenari acústic i l'SVM torna a col·lapsar cap a la classe control.

Aquest fenomen evidencia que la simple suma exhaustiva de descriptors sense netejar introdueix col·linearitat i soroll estadístic. Aquesta acumulació de dades redundants acaba confonent els algorismes durant la fase d'aprenentatge, fet que confirma la necessitat de dissenyar un procés de selecció de característiques a mida com el que s'aplica a la Configuració D.

9.2 Rendiment de la configuració optimitzada (D)

La **Configuració D** mostra els resultats obtinguts en el millor cas de la validació creuada després d'aplicar el filtratge de variables i la cerca d'hiperparàmetres. Com es pot veure a la Taula 5, la combinació d'aquestes tècniques fa créixer de manera notable la sensibilitat de tots els models.

TAULA 5
RESULTATS DE LA CONFIGURACIÓ D (MULTIMODAL OPTIMITZAT)

Configuració D				
Model	Acc.	Prec.	Recall	F2-Score
Random Forest	0.840	0.374	0.733	0.605
XGBoost	0.826	0.377	0.800	0.642
SVM	0.888	0.500	0.667	0.608

Nota de configuració (Hiperparàmetres òptims):

- **Random Forest:** 100 arbres, profunditat màxima de 3.
- **XGBoost:** 150 arbres, aprenentatge a 0.01, profunditat 5, submostreig al 70% i regularització L1 a 0.5.
- **SVM:** Kernel lineal, penalització $C = 0.005$ i selecció interna de 12 variables.

Per analitzar la capacitat predictiva global més enllà dels líndars fixats, s'avalua la corba ROC (Figura 7). Els valors d'Àrea Sota la Corba (AUC) resulten molt homogenis i elevats en les tres arquitectures, fet que demostra que tots els models assoleixen un potencial de separació de classes molt similar una vegada extret el soroll del dataset. No obstant això, l'estudi de les matrius de confusió (Figura 8) revela diferències crítiques en la distribució dels errors.

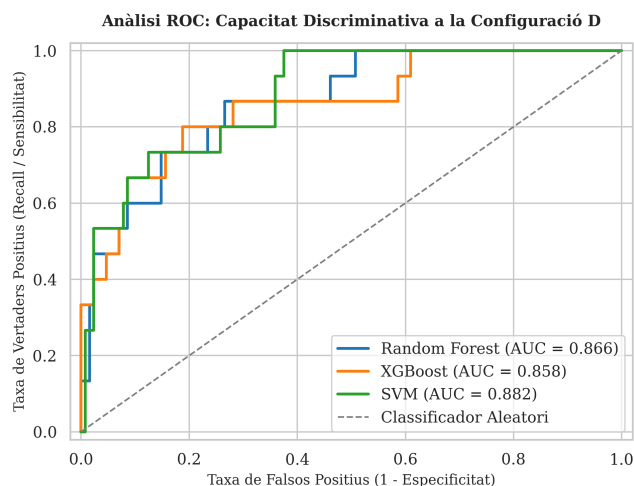


Fig. 7. Corbes ROC i valors AUC equivalents per als models optimitzats en la Configuració D.

		Random Forest (Configuració D)		XGBoost (Configuració D)		SVM (Configuració D)	
Classe Real	Control	109	19	108	20	117	11
	Risc	4	11	5	10	5	10
		Classe Predita		Classe Predita		Classe Predita	
		Control	Risc	Control	Risc	Control	Risc

Fig. 8. Matrius de confusió comparatives per als models Random Forest, XGBoost i SVM en la Configuració D.

El model **Random Forest** fa un salt important en la seva sensibilitat respecte a les fases anteriors, assolint un Recall del 73,3%. La seva matriu de confusió reflecteix aquest bon comportament registrant 11 encerts en la classe de risc i només 4 falsos negatius (errors d'omissió). Això s'explica perquè l'algorisme ha optat per arbres molt curts i simplificats. En limitar el creixement de les branques, el model perd una mica de precisió general (generant 19 falsos positius) però guanya molta capacitat per detectar els casos crítics, complint amb l'objectiu de cribratge preventiu de l'estudi.

Per la seva banda, l'**XGBoost** aconsegueix els millors resultats d'aquest escenari, registrant la sensibilitat més alta de la investigació amb un Recall del 80,0% i un F2-Score de 0,642. La seva matriu de confusió confirma aquesta eficàcia en cometre únicament 3 errors d'omissió davant de 12 infants detectats correctament com a risc. Per arribar a aquest rendiment, el model ha requerit un aprenentatge molt lent i controlat combinat amb l'ús d'una fracció reduïda de les dades per a cada arbre. Aquesta configuració evita que l'algorisme memoritzi el soroll i fa que es quedi només amb els patrons de veu i text més estables.

Finalment, la millora més significativa es veu en la **SVM**, que es recupera del col·lapse dels escenaris anteriors i registra l'atenuació de falsos positius més alta del bloc (només 11 casos, classificant correctament 117 infants

control). La seva matriu, de tota manera, reflecteix que aquesta alta precisió (50,0%) penalitza el Recall, ja que comet 5 falsos negatius (detectant 10 infants de risc). L'elecció d'un marge lineal i una penalització d'errors molt baixa indica que, un cop eliminada la informació redundant, el model prioritza un límit de decisió net i flexible, optimitzant la relació global tot i quedar per darrere dels arbres en la capacitat de detecció primerenca.

És necessari posar en relleu la baixa precisió relativa que registren els models d'arbres (*Random Forest*: 37,4%; *XGBoost*: 37,7%). Des d'una perspectiva purament matemàtica, un volum elevat de falsos positius es podria interpretar com una mancança de l'algorisme; no obstant això, en l'àmbit del cribratge psicopedagògic, aquest comportament respon exactament a la intenció d'aquest estudi.

En un filtre escolar preventiu, l'objectiu prioritari és maximitzar la sensibilitat (*Recall*) per reduir a zero els errors d'omissió (falsos negatius), ja que deixar un infant amb dificultats reals fora del radar té un cost educatiu irreversible. Per contra, un fals positiu només implica una posterior revisió individualitzada per part de l'especialista del centre per confirmar o descartar el cas.

Molts d'aquests falsos positius representen, en realitat, infants amb perfils de risc latent o dislèxies compensades que es beneficien directament d'aquesta detecció primerenca. Per poder visualitzar la naturalesa d'aquests subjectes, la seva proximitat a les fronteres de decisió i la distribució general de l'error, s'ha habilitat l'**Apèndix 4**, on s'analitza la distribució de probabilitats posteriors i índexs de distància a través de gràfics de dispersió interactius.

10 CONCLUSIÓ

El present treball ha assolit l'objectiu general de dissenyar i validar un sistema automatitzat i no invasiu capaç d'identificar de manera primerenca el risc de dislèxia a les escoles. Mitjançant l'ús de tècniques avançades d'aprenentatge automàtic aplicades a gravacions de lectures en veu alta, s'ha demostrat la viabilitat de transformar l'àudio en un predictor sòlid i accessible per a l'entorn educatiu.

Els resultats obtinguts són prometedors, ja que confirmen que la clau per a una detecció precisa es troba en l'enfocament multimodal. L'èxit del sistema es basa en la combinació de mètriques textuais i de fluïdesa lectora amb els descriptors acústics profunds de la veu de l'infant.

L'estratègia incremental dissenyada ha demostrat ser una decisió metodològica encertada per guiar el procés des d'uns models de referència inicials afectats pel fort desequilibri de la mostra —que tendien a ignorar el grup de risc— fins a unes configuracions finalment optimitzades. Aquesta progressió s'ha assolit gràcies a una evolució seqüencial basada en el filtratge exhaustiu

de les dades redundants i en la sintonització fina dels algorismes, aconseguint corregir la manca de sensibilitat inicial dels models i assolir el seu màxim potencial predictiu.

En definitiva, s'han aconseguit establir les bases d'una eina de cribratge preventiu robusta que prioritza la minimització dels falsos negatius. Això permet que la major part dels infants amb dificultats reals puguin ser detectats a temps, facilitant una intervenció primerenca dels especialistes del centre i posant de manifest el potencial de la intel·ligència artificial com a eina de suport preventiu a les aules.

Finalment, les línies futures del projecte es focalitzen en l'obtenció d'una mostra d'àudios més àmplia i representativa. L'accés a un volum de dades més gran permetrà reajustar i refinar els experiments desenvolupats, assegurant que el sistema guanyi precisió i es consolidi com un recurs de cribratge altament fiable per al sistema educatiu.

ÚS DE LA IA GENERATIVA

En el desenvolupament d'aquest projecte s'han utilitzat eines d'intel·ligència artificial generativa com a suport i assistència metodològica. Concretament, la IA s'ha emprat com a recurs d'optimització lingüística en la redacció i refinament formal tant dels informes corresponents a les entregues parcials com de la memòria final d'aquest article.

Així mateix, l'eina ha servit d'ajuda per estructurar el document i com a suport tècnic tant per dissenyar i maquetar els gràfics i figures resultants de l'estudi, com per millorar l'estructura del codi desenvolupat, afegint-hi comentaris explicatius i assegurant que fos el més net, llegible i entenedor possible. En tots els casos, la responsabilitat de les idees, el disseny experimental i les conclusions obtingudes és íntegrament de l'autor.

AGRAÏMENTS

En primer lloc, vull expressar el meu sincer agraïment a la meua tutora, Marta Prim Sabrià, per haver-me ofert total llibertat a l'hora de plantejar i desenvolupar els experiments d'aquest projecte. Li agraeixo profundament la seva constant disposició per guiar-me, així com la seva gran implicació i esforç a l'hora de contactar amb els centres educatius per gestionar l'obtenció de les dades necessàries.

Així mateix, vull manifestar el meu agraïment a les escoles col·laboradores, com també a les famílies dels participants. La seva valuosa confiança ha estat fonamental per permetre la recollida dels enregistraments d'àudio, un material essencial sense el qual aquest treball no hauria estat possible.

BIBLIOGRAFIA

- [1] XTEC – Àgora, “Com observar i detectar la dislèxia en els infants,” Informació sobre característiques i signes per identificar dificultats lectores i dislèxia, 2026. [En línia]. Disponible a: <https://agora.xtec.cat/escsolcunit/general/primaria/especialitats/educacio-especial/com-observar-i-detectar-la-dislexia-en-els-nens/>
- [2] ARA.cat, “Educació farà cribratges a tots els alumnes per detectar problemes d'aprenentatge,” Article sobre la iniciativa del Departament d'Educació per implementar cribratges escolars, 2026. [En línia]. Disponible a: https://www.ara.cat/societat/educacio/educacio-fara-cribratges-tot-s-alumnes-detectar-problemes-d-aprenentatge_1_5310474.html
- [3] J. Radford, G. Richard, H. Richard, i M. Serrurier, “Detecting Dyslexia from Audio Records: An AI Approach,” *Scitepress – Science and Technology Publications*, 2021. [En línia]. Disponible a: <https://www.scitepress.org/Papers/2021/101960/101960.pdf>
- [4] M. F. Rahman, R. Darni, D. Novaliendry, i K. Budayawan, “An Artificial Intelligence-Based Mobile Application for Early Detection of Dyslexia Using Recurrent Neural Network,” *Journal of Hypermedia & Technology-Enhanced Learning*, vol. 4, no. 1, pp. 49–67, 2026. [En línia]. Disponible a: <https://edutech-journals.org/index.php/j-hytel/article/view/217/107>
- [5] M. Sadiq et al., “A Review of Artificial Intelligence-Based Dyslexia Detection Techniques,” *Diagnostics*, vol. 14, no. 21, p. 2362, 2024. [En línia]. Disponible a: <https://www.mdpi.com/2075-4418/14/21/2362>
- [6] Projecte AINA, “Models de parla – Projecte AINA,” Documentació dels models de parla en català (Whisper i HuBERT), *BSC GitBook*, 2026. [En línia]. Disponible a: <https://langtech-bsc.gitbook.io/aina-kit/models/models-de-parla>
- [7] Milvus AI Reference, “Why is mean pooling often used on the token outputs of a transformer like BERT to produce a sentence embedding,” Explicació del mètode d'agregació d'embeddings, 2026. [En línia]. Disponible a: <https://milvus.io/ai-quick-reference/why-is-mean-pooling-often-used-on-the-token-outputs-of-a-transformer-like-bert-to-produce-a-sentence-embedding>
- [8] M. Barański, J. Jasiński, J. Bartolewska, S. Kacprzak, M. Witkowski, i K. Kowalczyk, “Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio,” *Signal Processing Group, Institute of Electronics, AGH University of Krakow*, preprint arxiv:2501.11378v1, 2025. [En línia]. Disponible a: <https://arxiv.org/pdf/2501.11378v1>
- [9] Healthcare Brew, “OpenAI’s transcription tool Whisper makes up words patients have never said,” Article analític sobre el problema de les al·lucinacions en entorns crítics, 2024. [En línia]. Disponible a: <https://www.healthcare-brew.com/stories/2024/11/18/openai-transcription-tool-whisper-hallucinations>
- [10] Python Software Foundation, “difflib — Helpers for computing deltas,” Documentació oficial sobre el funcionament de l'algorisme SequenceMatcher, 2026. [En línia]. Disponible a: <https://docs.python.org/3/library/difflib.html>
- [11] Scikit-Learn Developers, “Principal Component Analysis (PCA),” Documentació tècnica sobre la reducció de dimensionalitat mitjançant descomposició SVD, 2026. [En línia]. Disponible a: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

APÈNDIX

A1. PLANIFICACIÓ

La planificació del projecte s'ha estructurat de manera seqüencial al llarg del calendari acadèmic, dividida en les següents fases:

Setmanes	Fase de treball / Bloc executiu	Tasques clau i lliuraments realitzats
S5 - S6	Preparació i Preprocessament	Avaluació inicial del conjunt d'àudios per descartar gravacions tallades o intel·ligibles. Normalització dels audios.
S7 - S9	Extracció i Construcció del Dataset	Extracció de mètriques lectores. Generació d'embeddings acústics profunds amb <i>HuBERT</i> . Lliurament de l'Informe de Progrés I (S9).
S10 - S12	Desenvolupament i Entrenament	Implementació inicial dels classificadors <i>SVM</i> , <i>Random Forest</i> i <i>XGBoost</i> . Proves basades en 4 escenaris experimentals (textual, acústic, multimodal complet i optimitzat).
S13 - S15	Anàlisi Comparativa i Optimització	Comparació del rendiment dels models basant-se en les mètriques de recall i F2-score, i ajust d'hiperparàmetres per optimitzar la detecció. Lliurament de l'Informe de Progrés II (S14).
S16 - S17	Redacció i Tancament	Redacció de la memòria final de l'article. Lliurament de la proposta d'Informe Final (S17).
S18 - S19	Preparació de la Defensa	Elaboració de les diapositives de suport visual i assajos finals abans de la defensa del TFG. Lliurament de la Proposta de Presentació (S19).

A2. TEXT DE REFERÈNCIA I PONDERACIÓ DE LA COMPLEXITAT LÈXICA

En Carles

En Carles volia anar al cinema [2] amb els seus amics, però els seus pares no el deixaven [3]. Molt empipat [3], va entrar [2] a la seva habitació [3], va obrir [2] la guardiola [4] on guardava [3] els estalvis [3] i en va treure [3] unes quantes [2] monedes [2]. Durant [2] uns moments [3] va pensar [2] a baixar [3] per la finestra [3], però els seus pares s'enfadarien [5] molt, així [3] que no ho va fer. Va agafar [3] el telèfon [2] i va trucar [2] als seus amics que l'estaven [5] esperant [3]. Després [2] va estar [2] una estona [2] estirat [3] al llit fins que li va passar [2] l'empipament [5] i, llavors, ja més alegre [3], va anar a mirar la televisió [4] amb els seus pares.

A3. Matriu de correlació de Pearson

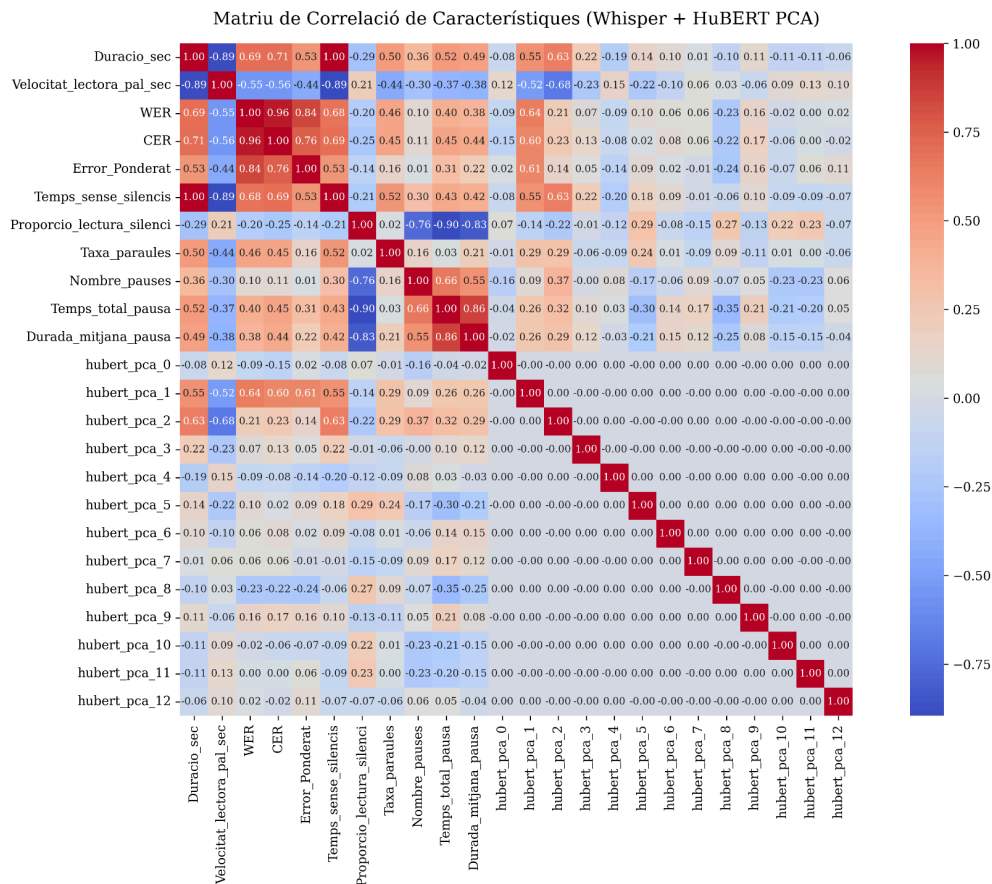


Fig. 8. Matriu de correlació de Pearson per a les variables candidates de Whisper i HuBERT PCA.

A4. Gràfics de Dispersió de la Incertesa Diagnòstica

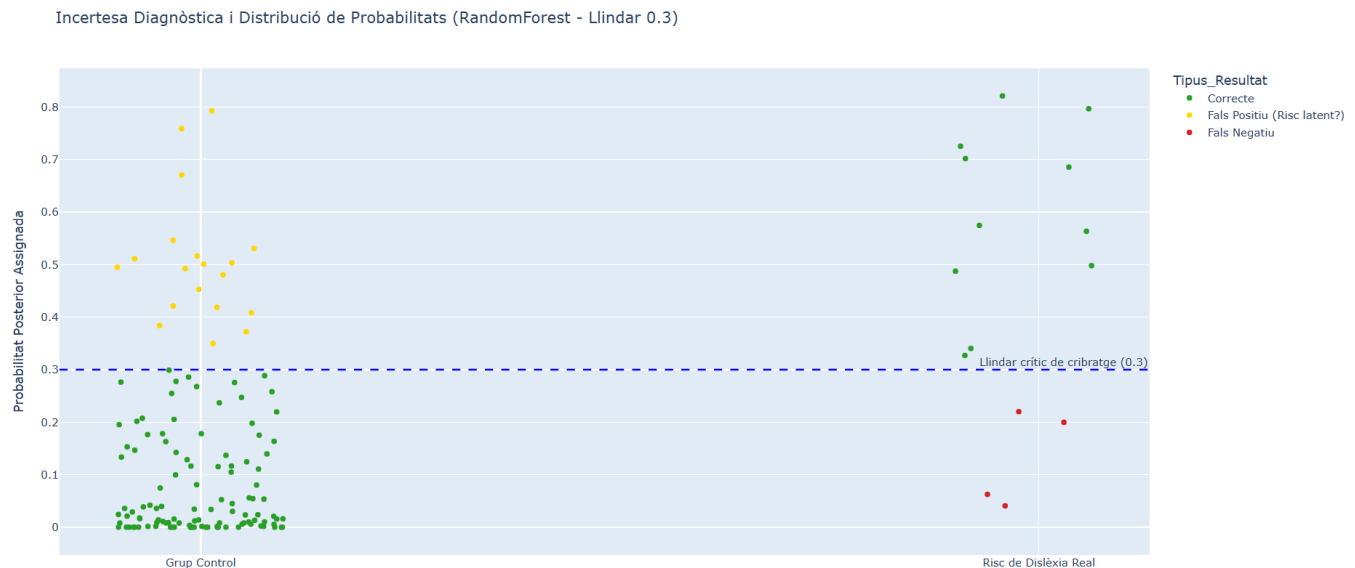


Fig. 9. Distribució de probabilitats posteriors per al model RandomForest (Configuració D guanyadora).

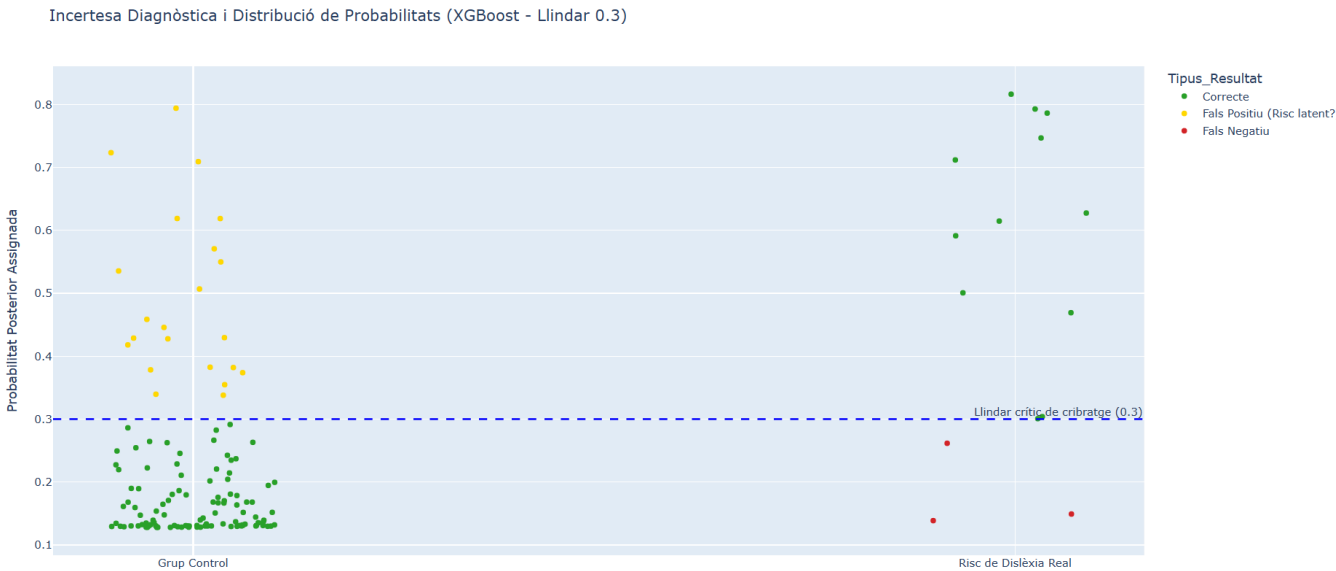


Fig. 10. Distribució de probabilitats posteriors per al model XGBoost (Configuració D guanyadora).



Fig. 11. Distribució de l'incertesa diagnòstica per al model SVM Lineal (Configuració D guanyadora). L'eix vertical mostra les distàncies geomètriques pures a l'hiperplà original normalitzades linealment en un rang [0, 1] per homologar la seva lectura amb la resta de models; el valor 0.5 actua com a límit net de decisió harmònica.