Desenvolupament d'un sistema de reconeixement de Llengua de Signes

Josep Maria Torres Ruiz

Resum— En aquest treball s'estudia el reconeixement automàtic de llengua de signes mitjançant tècniques de visió per computador i aprenentatge profund. Es presenten diverses arquitectures que combinen xarxes convolucionals (CNN) amb models seqüencials (LSTM i Transformers), incorporant mecanismes d'atenció per millorar la comprensió temporal. També s'exploren estratègies de tractament de dades, detecció de regions rellevants amb YOLOv8 i solucions al desbalanceig de classes. El projecte proporciona una anàlisi comparativa de diferents enfocaments i proposa línies futures per millorar el reconeixement de signes en vídeo.

Paraules clau— Llengua de signes, reconeixement automàtic, visió per computador, deep learning, CNN, LSTM, transformers, atenció, YOLOv8.

Abstract— This project explores the task of automatic sign language recognition using computer vision and deep learning techniques. Several architectures are proposed, combining convolutional neural networks (CNNs) with sequential models such as LSTM and Transformers, enhanced with attention mechanisms to better capture temporal dynamics. The study also includes data preprocessing strategies, key region detection using YOLOv8, and methods to address class imbalance. The project offers a comparative analysis of different approaches and outlines future directions for improving sign recognition from video sequences.

Index Terms— Sign language, automatic recognition, computer vision, deep learning, CNN, LSTM, transformers, attention, YoloV8.

1 Introducció

Actualment, s'estima que entre 70 i 300 milions de persones a tot el món utilitzen la llengua de signes com a mitjà principal per comunicar-se[1]. Aquesta xifra engloba tant persones sordes o amb discapacitat auditiva com aquelles que hi conviuen o treballen estretament, com familiars, intèrprets o professionals de l'àmbit educatiu i sanitari. Només a Espanya, es calcula que hi ha prop de 150.000 usuaris habituals de llengua de signes[2], una dada que reflecteix la importància i la necessitat d'aquest sistema comunicatiu dins la societat.

Malgrat el gran nombre de persones que depenen de la llengua de signes per a les seves interaccions quotidianes, cal destacar que aquesta no és universal. Existeixen múltiples llengües de signes arreu del món, cadascuna amb la seva pròpia gramàtica, lèxic i estructures comunicatives [3]. Aquestes diferències responen a factors culturals, històrics i geogràfics, de manera similar al que passa amb les llengües orals.

La llengua de signes és un sistema de comunicació visual-gestual que es fonamenta en l'ús de les mans, els braços, les expressions facials i altres moviments corporals per transmetre informació. Aquest sistema permet que les persones sordes o amb pèrdua auditiva puguin expressar idees complexes i comprendre els altres amb fluïdesa , ajudant a la qualitat de la comunicació.

En el marc d'aquest projecte, s'ha optat per treballar amb l'ASL (American Sign Language), considerada una de les llengües de signes més estandarditzades i difoses a escala internacional. L'ASL no només és àmpliament utilitzada als Estats Units i al Canadà, sinó que la seva influència ha arribat a diversos països, on ha estat adoptada o ha contribuït al desenvolupament de llengües de signes locals. La seva estructura gramatical ben documentada i l'extensa disponibilitat de recursos la converteixen en una elecció idònia per a projectes de reconeixement automàtic.

En els darrers anys, el reconeixement automàtic de la llengua de signes ha experimentat una evolució notable, especialment entre el 2015 i el 2020, període en què s'ha produït un increment de pocs estudis a més de 300[4]. Aquest creixement ha anat acompanyat d'un canvi de

paradigma. S'ha passat de sistemes intrusius basats en guants, sensors o càmeres especialitzades a solucions no intrusives fonamentades en visió per computador i càmeres RGB convencionals.

A més els avenços recents en reconeixement de signes s'han impulsat mitjançant tècniques de deep learning com les CNN, LSTM i, més recentment, els Transformers, que permeten capturar dependències temporals complexes. Els models híbrids com CNN-LSTM han demostrat bons resultats en entorns controlats amb vocabulari limitat. No obstant això, el camp encara presenta reptes, com la poca integració de paràmetres no manuals (expressions facials, mirada) i la dependència de corpus específics com RWTH-PHOENIX-Weather[4]., fet que dificulta la generalització i l'estandarditzación dels sistemes.

Sabent això en aquest informe, es presenta la implementació d'un sistema de reconeixement automàtic de signes mitjançant deep learning i visió per computador, amb l'objectiu de reconèixer signes realitzats en vídeo. Aquest treball vol contribuir a l'avenç de sistemes inclusius i accessibles que puguin facilitar la comunicació entre persones sordes i oients, establint una base sòlida per a futurs desenvolupaments.

2. Objectius

Aquest treball té com a finalitat el desenvolupament d'un model de reconeixement de la Llengua de Signes Americana (ASL) mitjançant tècniques de *deep learning* aplicades al processament de vídeo. Es plantegen els següents punts:

- **Traducció de vídeos a llenguatge escrit**: Crear un model capaç d'identificar i transcriure signes presents en seqüències temporals de vídeo de manera precisa i eficient.
- **Processament multimodal**: Aplicar tècniques de visió per computador per detectar i seguir els moviments manuals i facials, extraient-ne les característiques més rellevants.
- **Anàlisi temporal i contextual**: Implementar un enfocament seqüencial que permeti segmentar i classificar els signes dins d'un context temporal coherent, superant la interpretació aïllada.
- **Implementació en temps real**: Aconseguir un sistema funcional capaç de realitzar prediccions amb latència reduïda, adequat per aplicacions pràctiques i en temps real.

3 Metodologia

Per al desenvolupament del projecte, es planteja una metodologia que combina visió per computador i deep learning. Aquest enfocament integrat proporciona un marc de treball robust i adequat per abordar de manera eficient els objectius del projecte i garantir-ne una implementació completa i precisa.

3.1 Datasets

Inicialment, per a la resolució del projecte, és necessari disposar de dades d'entrenament. En aquest cas, tal com s'explicarà amb més detall en l'apartat de metodologia, es faran servir dos conjunts de dades. El primer és el WLASL [5][6], que s'utilitzarà per entrenar el model encarregat de traduir vídeos a paraules lèxiques. El segon és el hand_face_detection_dataset [7], que servirà per entrenar un model YOLO destinat a ser utilitzat durant el procés de tractament de dades.

La selecció d'aquests conjunts de dades s'ha fet tenint en compte la seva adequació a l'aplicació plantejada per a la resolució del problema, així com la seva facilitat d'ús. A més, ofereixen una gran diversitat de contextos visuals, fet que afavoreix una millor variabilitat durant l'aprenentatge del model.

3.1.1 WLASL

Per a l'entrenament s'ha utilitzat el conjunt de dades WLASL [5][6] (Word-Level American Sign Language), un dels recursos més complets disponibles per al reconeixement de signes a nivell lèxic en llengua de signes americana. El dataset inclou 12.000 vídeos associats a prop de 2.000 glosses diferents.

Els vídeos provenen de fonts públiques i presenten una elevada variabilitat en qualitat d'imatge, condicions d'il·luminació, colorimetria i presència de reflexos. Malgrat això, tots estan enregistrats a 25 fotogrames per segon (fps), fet que assegura una uniformitat temporal mínima.

Un dels punts forts del WLASL és la diversitat de signants, que introdueix variabilitat significativa en la representació dels signes (velocitat, estil, precisió). Aquesta heterogeneïtat suposa un repte per a l'aprenentatge automàtic, però també potencia la diferencia entre mostres.

Donada la seva naturalesa, el dataset ha requerit un procés de pretractament que inclou la selecció de vídeos vàlids, la normalització de fotogrames, l'homogeneïtzació de resolucions i, si escau, tècniques d'augmentació per millorar la qualitat de les dades d'entrada.

3.1.2 hand_face_detection_dataset

Per complementar el sistema de reconeixement, s'ha utilitzat el conjunt de dades Hand and Face Detection Focused on Sign Language [7], disponible a Kaggle. Aquest dataset està orientat a la detecció de mans i cares, elements clau en el processament de Llengua de Signes.

Inclou prop de 21.000 imatges anotades amb coordenades de bounding boxes per a mans i cares, fet que permet entrenar models de detecció amb localització precisa. Les imatges cobreixen una àmplia variabilitat en il·luminació,

postures i expressions facials, afavorint la capacitat de generalització del model en condicions reals.

Totes les imatges provenen de fonts públiques i estan específicament relacionades amb escenes de Llenguatge de Signes, fet que incrementa la rellevància del dataset per al domini

3.2 Tractament de les Dades

Tal com s'ha exposat a l'apartat 3.1.1, els vídeos del conjunt de dades presenten una gran variabilitat en termes de durada, resolució i qualitat. Per tal de garantir la coherència entre mostres i facilitar el processament posterior, es duu a terme un procés de normalització que permet adaptar els vídeos a un format uniforme i adequat per a l'entrada als models de deep learning.

En primer lloc, es redimensionen tots els vídeos a una resolució comuna de 224×224 píxels [8], una mida àmpliament utilitzada en arquitectures de visió per computador com VGGNet [9] o ResNet [10]. A més, és compatible amb molts models preentrenats disponibles en biblioteques com TensorFlow o PyTorch, els quals han estat entrenats sobre imatges d'aquesta mida.

Un cop establerta la resolució comuna de 224×224 píxels, cada vídeo es segmenta en fotogrames individuals per facilitar l'aplicació de tècniques de visió per computador. Per tal de centrar l'atenció del model en les regions més rellevants,mans i cara, s'utilitza l'arquitectura YOLOv8s, escollida per oferir una bona precisió amb un cost computacional moderat, millorant el rendiment en la detecció d'objectes petits en comparació amb YOLOv8n [11]. Aquest model es va entrenar durant 32 èpoques amb el conjunt del apartat 3.1.2, assolint així una precisió del 99,66 %, un recall del 99,52 % i una mAP@0.5 de 0,8867, amb una convergència clara a partir de l'època 15 i un temps d'inferència mitjà de 8,9 ms per fotograma en GPU. Un cop aplicat sobre els fotogrames, es duu a terme un postprocessament on només les regions detectades es mantenen a color, mentre que la resta de la imatge es transforma a negre, amb l'objectiu de reduir el soroll de fons i reforçar l'aprenentatge focalitzat del model.

Després del preprocessament visual, es normalitza la durada dels vídeos establint una longitud fixa de 110 fotogrames per seqüència. Aquesta decisió es fonamenta en l'anàlisi estadística del conjunt de dades, que revela una mitjana de 68,85 fotogrames per vídeo, un percentil 90 de 101 i un percentil 95 de 112. Establir aquest límit permet garantir l'homogeneïtat entre mostres i evita el processament de vídeos atípics que podrien introduir soroll o biaixos en l'aprenentatge. Per assolir aquesta uniformitat temporal, s'apliquen dues estratègies en funció de la durada original: si el vídeo supera els 110 fotogrames, se selecciona un subconjunt intermedi distribuït de manera uniforme al llarg de la seqüència; si en té menys, es dupliquen fotogrames intermedis fins a assolir la longitud requerida. Aquesta tècnica substitueix el padding tradicional ,basat en la repetició dels últims fotogrames, ja que, segons s'ha observat experimentalment, pot dificultar l'aprenentatge seqüencial. En concret, els resultats mostren una diferència clara a favor de la duplicació, amb una pèrdua de validació mitjana gairebé 0,8 punts inferior i una accuracy notablement més elevada des de les primeres èpoques.

Aquest conjunt de passos defineix un pipeline de preprocessament robust, eficient i coherent, que assegura que les seqüències introduïdes tinguin una estructura homogènia i representin amb claredat les característiques essencials del senyal visual per al reconeixement de signes.

3.3 Reducció de la dimensionalitat mitjançant xarxes convolucionals (CNN)

Un cop preprocesades les dades, cada vídeo es representa com una seqüència de 110 fotogrames RGB de mida 224×224, centrats en les regions d'interès (mans i cara). A causa de la seva elevada dimensionalitat $R^{3x224x224}$ per fotograma), es fa necessari aplicar una reducció per tal d'obtenir representacions més compactes i discriminatives. Per això, s'utilitza una xarxa neuronal convolucional ResNet18[12] com a extractor de característiques, aprofitant les seves quatre etapes de convolució principals. En concret, la sortida del bloc final (abans de la capa de classificació) genera un vector de 512 dimensions per imatge, que conté informació semàntica i estructural d'alt nivell.

Formalment, si denotem $x \in R^{3X224X224}$ en el fotograma en l'instant t, aquest és processat pel mòdul convolucional $CNN(\cdot)$ per obtenir la seva representació vectorial:

$ft=CNN(xt)$, $ft \in R^{512}$

Amb això, una seqüència de T=110 fotogrames es transforma en una matriu de característiques:

$F=[f1,f2,\ldots,fT] \in R^{110X512}$

Per tal d'adaptar aquesta matriu al model seqüencial (LSTM o Transformer), es projecta cada vector ft a una nova dimensionalitat dmodel=256 mitjançant una transformació lineal:

$zt=Wft+b$, $zt \in R^{256}$

on $W \in R^{256X512}$ és la matriu de pesos i $b \in R^{256}$ és el vector de biaix. Aquesta projecció redueix la dimensionalitat i facilita l'entrenament del model seqüencial. Finalment, s'hi afegeix una codificació posicional per conservar la informació temporal abans d'introduir les representacions al model (ja sigui LSTM o Transformer), que s'encarrega de capturar les dependències temporals entre fotogrames.

3.4 Mecanismes d'atenció aplicats a seqüències

Amb l'objectiu d'optimitzar la representació contextual

d'una seqüència de descriptors espacials obtinguts mitjançant una CNN, s'incorporen mecanismes d'atenció com a capa d'agregació adaptable sobre les sortides del model seqüencial. Aquests mecanismes permeten relaxar la dependència exclusiva de l'última sortida del LSTM i facilitar l'extracció d'informació rellevant distribuïda al llarg de la seqüència.

S'han implementat tres variants diferenciades:

1. **Atenció de Bahdanau (additive attention):** aquest mecanisme, introduït en el context de traducció automàtica neuronal [13], aplica una transformació no lineal sobre cada pas temporal per calcular-ne la rellevància. Les puntuacions et=v^T tanh(W1ht+W2s+b) després es normalitzen mitjançant *softmax*, i el vector de context s'obté com a combinació ponderada de les sortides del LSTM.

2. **Multi-Head Self-Attention (MHSA):** proposta central en l'arquitectura Transformer [14], aquest mecanisme permet que cada pas temporal atengui tots els altres mitjançant múltiples projeccions paral·leles. Quan s'utilitza com a capa final d'una xarxa seqüencial, ofereix una rica capacitat per capturar dependències globals.

3. **Atenció temporal lineal:** implementada mitjançant una capa et=Wht amb W∈R1×D que assigna pesos escalars per cada ht∈RD, aquesta versió més lleugera computacionalment ha estat efectiva en tasques seqüencials amb longitud moderada [15].

Totes les variants generen un vector de context c∈R^D que resumeix adaptativament la seqüència i millora la sensibilitat del model als instants temporals més informatius.

3.5 Modelització de la seqüència: LSTM vs Transformer

Per capturar la dependència temporal entre fotogrames successius, s'han considerat dues arquitectures seqüencials: una basada en xarxes LSTM i una altra fonamentada en el paradigma Transformer. Ambdues processen com a entrada una seqüència X=[x1,…,xT] de vectores de característicasa $xt∈$R^{512}$,obtinguts a partir del mòdul convolucional.

3.5.1 Xarxa LSTM amb atenció

Les xarxes Long Short-Term Memory (LSTM) són una extensió de les RNN convencionals que mitiguen el problema del gradient esvaït mitjançant mecanismes interns de control com les portes d'entrada, sortida i oblit [16]. Aquestes estructures permeten preservar informació rellevant durant períodes temporals llargs, sent útils en tasques seqüencials com el reconeixement de signes. En aquest projecte s'han avaluat dues variants:

3.5.1.1 LSTM unidireccional

Processa la seqüència d'entrada X=[x1,…,xT] de forma cronològica. Cada estat ocult h_t depèn exclusivament dels vectors anteriors, és a dir, ht=LSTM(xt,ht−1). Aquest enfocament és útil quan cal respectar l'ordre temporal estricte i la causalitat.

3.5.1.2 LSTM bidireccional

Incorpora dues LSTM independents: una que processa la seqüència en sentit directe i una altra en sentit invers. La sortida final de cada pas temporal és la concatenació de les dues direccions:

$$ht=[htd;hti]∈R^{2dh}$$

Això permet capturar dependències tant passades com futures dins la seqüència, fet que pot millorar la representació quan no cal respectar la causalitat estricta.

Per a cada vídeo, la sortida de la xarxa LSTM és una matriu H∈R^{TXD}, on D = 512 en el cas bidireccional (256 per direcció). Aquesta matriu es pot processar de dues maneres: mitjançant una agregació estàtica (com el mean pooling) o per un mòdul d'atenció, que pot ser Bahdanau, Self-Attention o temporal lineal, els quals generen un vector de context $C = R^D$mitjançant una agregació adaptativa de les sortides temporals. Finalment, aquest vector s'introdueix a una capa lineal de classificació,y=softmax(Wc·c+bc). Aquesta operació dona com a resultat una distribució de probabilitat sobre les possibles classes, indicant la probabilitat que el vídeo pertanyi a cadascuna d'elles.

3.5.2 Transformer Encoder:

Les xarxes Transformer representen una arquitectura seqüencial alternativa que substitueix completament la recursivitat per mecanismes d'atenció autoaplicada (self-attention) i processament paral·lel [17].

En aquest projecte s'ha implementat un Transformer Encoder, que aplica múltiples capes d'atenció multi-cap (multi-head self-attention) i blocs feed-forward independents per a cada posició.

En aquest cas, la seqüència X es projecta a un espai de dimensió,dmodel=256,mitjançant una capa lineal, i s'enriqueix amb una codificació posicional sinusoidal P∈ $R^{TxDmodel}$ per obtenir Z=XW+b+P∈$R^{TxDmodel}$. Aquest tensor Z és processat per un Transformer Encoder. El vector global es calcula mitjançant mean pooling: c=T1t=1∑Tzt(L)∈R^{Dmodel} és la sortida de l'última capa per la posició t.

4 EXPERIMENTS

Per la obtenció de el model resultant aplicant aquesta metodologia [A.1][A.2], es duran a terme diversos experiments amb l'objectiu d'obtenir una solució raonada i ajustada al problema plantejat.

4.1 Transformers VS LSTM

En aquest primera part del estudi s'han desenvolupat i comparat dues arquitectures de deep learning per al reconeixement automàtic de llengua de signes. Totes dues comparteixen una xarxa convolucional ResNet18 com a extractor de característiques espacials, però difereixen en l'enfocament temporal.

La primera arquitectura combina la ResNet18, amb les seves capes congelades, amb una capa LSTM. La segona, en canvi, es basa en un codificador Transformer que incorpora codificació posicional sinusoidal i diverses capes d'atenció multi-cap per modelar relacions temporals globals dins la seqüència. Ambdues arquitectures s'han entrenat sota les mateixes condicions (2.000 classes i batch size 4), fet que permet comparar-ne els punts forts i febles de manera objectiva dins el context del projecte.

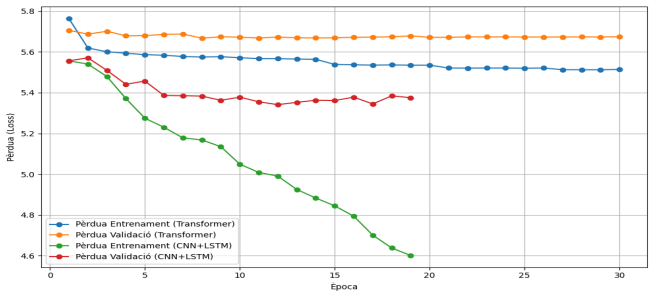


Fig 1. Comparativa de les corbes de *loss* entre les arquitectures CNN+Transformer i CNN+LSTM

Les corbes d'entrenament mostren que cap dels models aconsegueix un aprenentatge òptim. En el cas del Transformer, la manca de millora en train i test indica la necessitat d'aplicar early stopping. A més en ambdós casos, la CNN podria no estar capturant bé les característiques dels vídeos, possiblement per la variabilitat entre seqüències; per això, descongelar-la parcialment podria millorar la convergència.

Pel que fa al model amb LSTM, es podria detectar overfitting, ja que el rendiment en entrenament és superior al de validació. Incorporar dropout i utilitzar una BLSTM, com fan molts estudis [4], podria millorar el resultat. Finalment, la distribució desequilibrada de classes[A.3] dificulta l'aprenentatge; per això, es va optar per entrenar només amb les classes més freqüents.

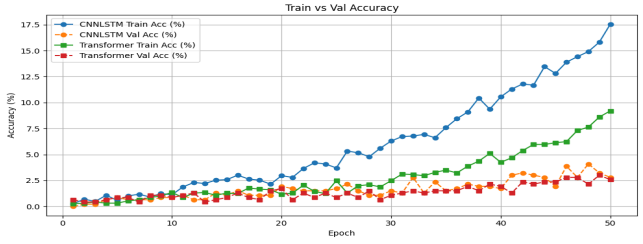


Fig 2. Comparativa de les corba d'accuraccy entre les arquitectures CNN+Transformer i CNN+BLSTM amb millores

L'anàlisi de la corba d'entrenament posa de manifest que les millores implementades han tingut un impacte positiu en el rendiment del model. En concret, la descongelació parcial de la CNN i la reducció del nombre de classes a les 256 amb més representació,amb un mínim de 15 mostres per classe,han afavorit una millor convergència i uns resultats globals més sòlids. Aquestes modificacions també han contribuït a una reducció més efectiva de la pèrdua durant l'entrenament [A.4].

En la comparació entre arquitectures, s'observa que els models amb BLSTM obtenen lleugerament millors resultats que els basats en Transformers, fet que justifica centrar la investigació en aquesta direcció. A més, si comparem les Figures 1 i 2, es constata una millora clara entre l'arquitectura LSTM inicial i la versió amb BLSTM i dropout, consolidant aquesta última com a proposta base per continuar amb el desenvolupament del sistema.

Tot i les millores aconseguides, encara existeixen línies clares de millora que es poden explorar en fases posteriors. En primer lloc, cal replantejar l'estratègia d'entrenament davant la presència d'un gran nombre de classes. Una alternativa és l'entrenament incremental per blocs de classes, on el model s'exposa progressivament a subconjunts reduïts de paraules, permetent una assimilació més controlada de la variabilitat lèxica. Aquesta estratègia té mecanismes de rehearsal per reforçar la retenció de les classes ja apreses i evitar el fenomen del catastrophic forgetting [18].

Seguidament, es consideren diverses millores rellevants tant a nivell d'optimització arquitectònica com de regularització per les següents proves. Una primera proposta és la congelació parcial de la CNN durant les primeres èpoques per estabilitzar l'inici de l'entrenament i preservar les representacions preentrenades [19]. A nivell seqüencial, es planteja combinar estratègies de pooling, com la mitjana temporal (mean pooling) i la sortida final de la LSTM, per obtenir una representació més robusta del senyal. També s'incorporen tècniques d'augment de dades com el flip horitzontal, el color jitter i el blur, que milloren la generalització i redueixen el risc d'overfitting, especialment en entorns amb dades escasses o desequilibrades. Finalment, el batch size s'ha mantingut en 4 per limitacions tècniques, fet que pot afectar l'estabilitat de l'entrenament i es considera un possible aspecte a optimitzar en treballs futurs.

4.2 Entrenament Incremental

En aquesta prova s'ha implementat una estratègia d'entrenament incremental per abordar la dificultat d'aprenentatge en entorns amb moltes classes i desequilibris en la distribució. El model utilitzat, basat en una arquitectura CNN+BLSTM amb dropout i augmentació espaciotemporal, es va entrenar de manera progressiva en fases, afegint blocs reduïts de classes en cada etapa. A cada fase, la capa de classificació s'adapta al nou conjunt de classes i es preserva el coneixement adquirit mitjançant la congelació parcial de la xarxa. Aquest enfocament pretén facilitar un aprenentatge seqüencial estructurat i reduir el risc de catastrophic

forgetting. No obstant això, els resultats obtinguts ,representats a la Figura 5, posen de manifest que el corpus no disposa d'una quantitat suficient de vídeos per classe. Aquesta escassetat provoca que, en dividir les classes en subconjunts, el model no només aprengui de manera limitada, sinó que també oblidi fàcilment el que ha après, com es reflecteix en les oscil·lacions pronunciades observades als registres de log.

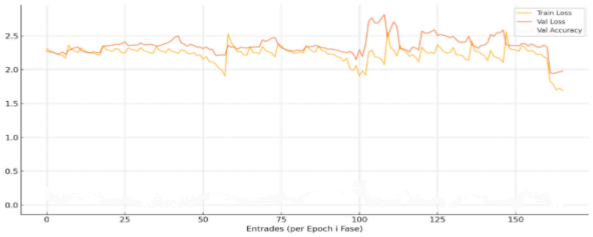


Fig 3. Resultats obtinguts amb l'estratègia d'entrenament incremental per blocs de classes.

L'anàlisi de la grafica (fig.3) permet descartar aquesta aproximació, ja que la quantitat de vídeos per classe és massa reduïda i la seva distribució és altament dispersa [A.3], fet que dificulta un aprenentatge efectiu en un esquema incremental.

4.3 Millores BLSTM

En aquesta millora del model s'ha implementat una arquitectura CNN+BLSTM amb estratègies de regularització per millorar la generalització. S'ha utilitzat una ResNet18 preentrenada amb congelació parcial durant les primeres èpoques per estabilitzar l'entrenament. La sortida de la CNN s'ha processat amb una LSTM bidireccional de dues capes, combinant el mean pooling i la darrera sortida temporal per capturar la dinàmica global i la informació final del senyal. A més, s'ha aplicat augmentació espaciotemporal sobre un subconjunt de fotogrames (flip horitzontal, color jitter i blur) per incrementar la robustesa visual, així com dropout i early stopping per afavorir l'estabilitat del model.

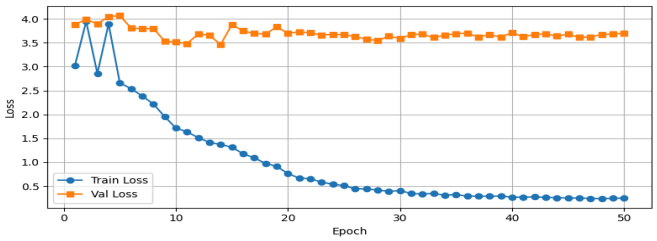


Fig 4. Resultats de la Loss obtinguts amb la millora del model CNN+BLSTM

Els resultats obtinguts evidencien una millora substancial respecte al model base, amb una reducció de la funció de pèrdua (Fig.4) i un increment de l'accuracy de validació fins al 22% [A.5]. No obstant això, l'anàlisi de les corbes d'aprenentatge revela un patró acusat d'overfitting, reflectit en una diferència persistent entre la loss d'entrenament i la de validació (0.5–3.5 unitats). Tot i aquest desajust, l'aplicació de tècniques com la congelació parcial de la CNN, el pooling híbrid i l'augment de dades ha permès una millor captació de patrons complexos. Tanmateix, la capacitat de generalització del model continua limitada pel desequilibri en la distribució de classes [20].

Per abordar aquestes limitacions, es proposen quatre línies de millora. L'aplicació de pesos inversament proporcionals a la freqüència de cada classe dins la funció de pèrdua pot compensar l'impacte de les classes majoritàries[21]. També, una reducció addicional del conjunt de classes en l'aprenentatge. A més l'ús de learning rate adaptatiu, mitjançant estratègies basades en el loss, pot millorar la convergència i evitar oscil·lacions en la validació [22]. Finalment, la incorporació de mecanismes d'atenció simples (Atenció lineal 3.4), podria permetre al model identificar i focalitzar-se en les regions més informatives de la seqüència, millorant així tant la representació com la decisió final [23].

4.4 Model SAFINet

Abans d'aplicar les millores al model CNN+BLSTM, es va implementar una versió funcional del model SAFINet (Spatial-Temporal Multi-Cue Attention Fusion Network) proposat per Zhou et al.[24]. Aquesta arquitectura jeràrquica captura patrons espaciotemporals en seqüències de vídeo mitjançant una ResNet18 com extractor espacial, seguida de dos mòduls Inception1D per capturar patrons temporals multiescala, i un Transformer Encoder amb 4 capes i 8 caps d'atenció per modelar dependències a llarg termini. La classificació es fa a partir de la darrera posició del Transformer. La implementació, adaptada al reconeixement discret de signes, es va entrenar sobre el mateix corpus per determinar si les limitacions eren atribuïbles al model o a les dades.Amb aquesta implementació obtenim la fig.5.

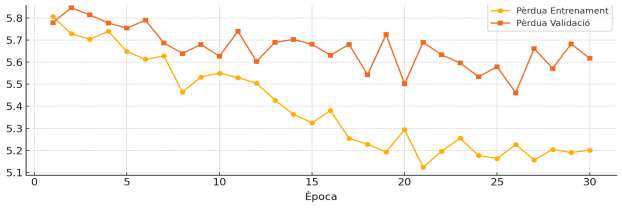


Fig 5. Resultats de la Loss SAFINet

Tal com reflecteix la corba d'entrenament, el model no ha aconseguit convergir, amb una evolució pràcticament plana. Aquesta manca d'aprenentatge pot ser conseqüència d'una inadequada adequació del corpus a l'arquitectura SAFINet, que requereix una major coherència temporal. A més, la variabilitat intra-classe i el desequilibri en la distribució de mostres dificulten la capacitat de generalització. Tot i aquestes limitacions, s'aplicaran les millores arquitectòniques i de regularització proposades per tal d'avaluar si poden compensar-les, especialment tenint en compte que, com

s'indica a l'apartat 4.3, l'esbiaix del dataset podria ser el que més negativament efectes.

4.5 Reducció de l'overfitting

Per començar a reduir overfitting., es va aplicar una reducció del nombre de classes amb l'objectiu d'avaluar si el baix rendiment del model es devia a una manca d'optimització per a escenaris amb moltes classes. Concretament, es van seleccionar únicament les classes que disposaven de més de 25 vídeos, donant com a resultat un subconjunt format per 8 classes en total.

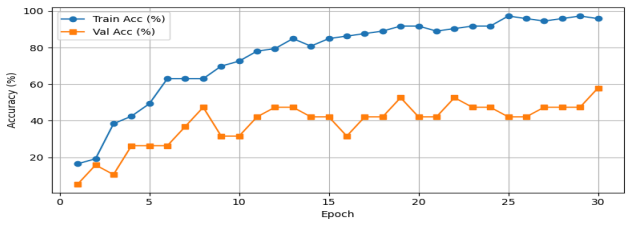


Fig 6. Resultat de l'accuracy obtingut amb la millora del model CNN+BLSTM amb 8 Clases

En observar la figura 6 , es pot apreciar que la reducció del nombre de classes comporta una millora en el rendiment quan el nombre de classes és més reduït. Tot i això, no s'aconsegueix eliminar del tot overfitting, cosa que suggereix que el model pot no estar generalitzant correctament les dependències temporals entre seqüències, probablement a causa de la variabilitat entre els vídeos. A més, si ens fixem en la matriu de confusió (figura 7), s'evidencien dificultats addicionals en la discriminació entre certes classes.

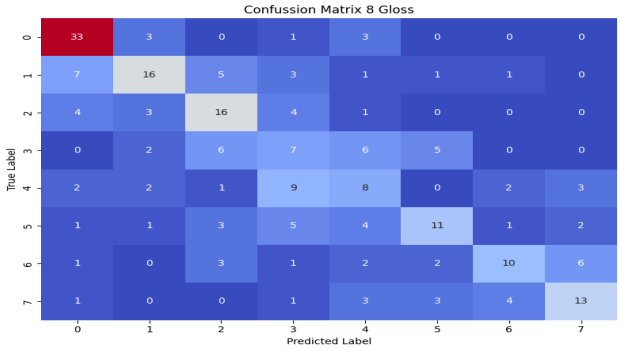


Fig 7. Matriu de confusió del model CNN+BLSTM amb 8 clases

S'observa clarament que les classes amb una major quantitat de vídeos també presenten una major representació en la matriu de confusió, amb valors més elevats a la diagonal, fet que indica una millor precisió en la classificació. Aquest fenomen posa de manifest el desbalanceig del conjunt de dades, ja que les classes amb menys mostres són més propenses a ser confoses amb altres categories. Això és perquè el model té més dificultats de generalitzar ja que en llenguatge de signes en cada mostra hi ha molta variabilitat.

Seguidament com es detalla a l'apartat 4..3, es va dur a terme les implementacions a una arquitectura CNN+BLSTM amb diverses millores específiques en optimització, regularització i balanceig de dades, amb l'objectiu d'avaluar el seu impacte sobre el rendiment. En concret, es va incorporar un mòdul d'atenció simple de tipus lineal per tal de capturar patrons temporals rellevants dins la seqüència de vídeo. A més, es va afegir un dropout a la sortida del mòdul d'atenció per afavorir la generalització, tal com s'havia definit a l'apartat 3.4.

Pel que fa a l'estratègia d'optimització, es va utilitzar l'optimitzador Adam amb weight decay, combinat amb un planificador adaptatiu de la taxa d'aprenentatge (ReduceLROnPlateau) que redueix automàticament la learning rate quan la validació s'estanca, millorant així l'estabilitat de l'entrenament. A més, es va aplicar early stopping per evitar el sobreajustament.

Finalment, per mitigar el desbalanceig de les classes, es van calcular pèsos inversament proporcionals a la seva freqüència, la qual cosa ajuda a compensar la manca de representació d'algunes glosses durant l'entrenament.

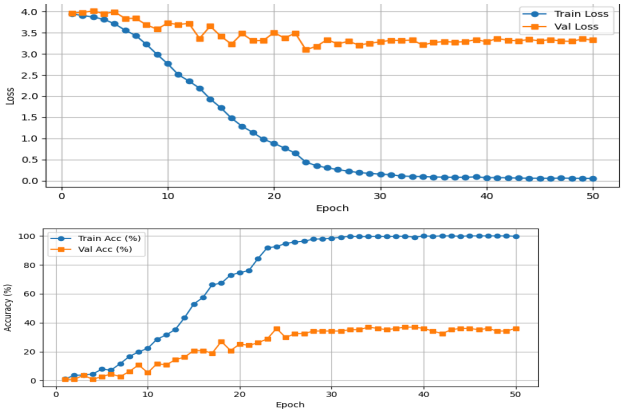


Fig 8. Resultats de la Loss i accuracy obtinguts amb la millora del model CNN+BLSTM+Attention i reguladors

Les grafiques (fig 8) mostren una clara millora en els resultats de loss i accuracy respecte a versions anteriors del model.Aquestes millores són especialment rellevants ara que s'està treballant amb 256 glosses, ja que la complexitat del problema augmenta substancialment respecte als experiments previs amb només 8 glosses.

Tot i aquestes millores, els resultats indiquen que el model encara no generalitza del tot bé, especialment en validació. Aquesta limitació podria estar relacionada amb la presència de desbalanceig en les dades, que no queda totalment resolta només amb regularització (weight decay) o càlcul de pesos per classe.

En aquest sentit, la incorporació de tècniques de reequilibri de dades, com l'oversampling de les classes minoritàries o el downsampling de les majoritàries, podria contribuir a millorar la capacitat de generalització del model.

A més, la introducció del mòdul d'atenció ja ha demostrat tenir un impacte positiu en els resultats (fig 8).Per tant, i tal com es proposa a l'apartat 3.4, implementarem el rendiment del model actual amb altres mecanismes d'atenció més sofisticats, com ara l'atenció de Bahdanau o la Multi-Head Attention, per avaluar si poden oferir una millora addicional en la captació de patrons temporals i semàntics complexos.

4.6. Proves attention

Per tal d'avaluar si les millores observades es poden potenciar encara més mitjançant diferents estratègies d'atenció, s'ha dut a terme una sèrie de proves comparatives amb dues arquitectures .

Aquestes dues versions del model s'han implementat de manera modular dins d'una arquitectura CNN+BLSTM, i s'ha mantingut la mateixa configuració base en tots els experiments per garantir una comparació justa. A més, s'han mantingut els mecanismes de regularització adaptativa com el dropout dinàmic (ajustat automàticament en funció de l'evolució de la validation loss) i el weight decay.

L'objectiu d'aquestes proves és identificar si aquests mecanismes d'atenció més sofisticats poden capturar millor les dependències temporals complexes entre fotogrames i oferir una millor representació contextual de les seqüències de vídeos. Aquesta anàlisi sistemàtica permet conèixer quin tipus de mòdul d'atenció contribueix més significativament a millorar la classificació en seqüències de vídeo de llengua de signes.

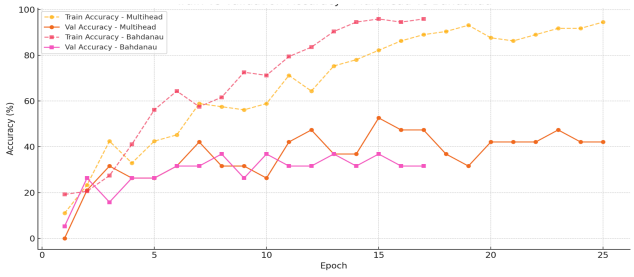


Fig 9. Comparativa de les corbes d'accuracy entre les arquitectures amb atenció Bahdanau i Multihead .

La comparativa entre els models amb atenció Multihead i Bahdanau, posa de manifest diferències significatives en el seu rendiment. El model amb atenció Multihead presenta resultats globals superiors, amb valors màxims de F1 per sobre de 0.5 i un Recall mitjà aproximat de 0.42, tot i mostrar una major variabilitat durant l'entrenament (fig 10). Per contra, el model amb atenció Bahdanau evoluciona de manera més estable, però amb valors mitjans més baixos (Precision ~0.32, Recall ~0.35). Així doncs, Multihead destaca com l'opció més eficient, especialment en tasques on el Recall és prioritari, com ara el reconeixement de llengua de signes. Aquesta anàlisi suggereix que, amb una arquitectura relativament senzilla, és possible assolir resultats prometedors en aquest àmbit.

4.7 Tècniques de reequilibri: Oversampling VS down sampling

Per altre banda, per tal de mitigar el desbalanceig de classes present al conjunt de dades, s'han aplicat dues estratègies de reequilibri.

L'oversampling consisteix a augmentar el nombre de mostres de les classes minoritàries mitjançant duplicació aleatòria, fent que totes les classes tinguin una representació similar. En canvi, el downsampling redueix el nombre de mostres de les classes majoritàries per igualar-les a les menys representades.Obtenit aixi una qunatitat balancejada pero diverse de mostres[A.6].

En la fig 10. Observem que en aquestt cas, els resultats amb l'estratègia d'oversampling mostren un comportament clarament limitat pel sobreajustament. L'accuracy d'entrenament assoleix valors molt elevats, superant el 95 % cap a la època 35, mentre que l'accuracy de validació es manté entorn del 27 % amb una desviació estàndard d'uns ±2.1 %, evidenciant una manca de generalització del model. Aquest patró també es reflecteix en les mètriques de classificació: la Precision arriba a una mitjana de 0.42 ± 0.05, mentre que el Recall i el F1 Score es mantenen al voltant de 0.33 ± 0.03. Aquests resultats indiquen que, en duplicar massivament vídeos per equilibrar les classes, s'ha reduït la diversitat del conjunt d'entrenament, i com a conseqüència el model aprèn patrons específics que no trasllada correctament a dades noves. Així doncs, tot i millorar el balanç de classes, l'oversampling en aquest cas ha compromès la capacitat del model per generalitzar, afectant negativament la precisió i la capacitat de detecció global.

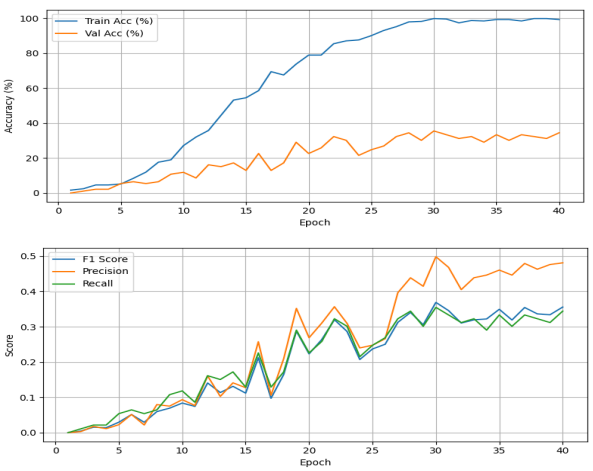


Fig 10. Corbes d'accuracy, precision, recall i F1 score obtingudes amb l'estratègia d'oversampling durant l'entrenament.

Els resultats obtinguts amb l'estratègia de downsampling (Fig.11) mostren una lleugera millora respecte a l'oversampling, especialment pel que fa a l'accuracy de validació, que en aquest cas arriba fins a un 40 % ± 2.4 %. Així mateix, les mètriques de Precision, Recall i F1 Score també augmenten lleugerament, amb valors mitjans al

voltant de 0.44 ± 0.06, 0.35 ± 0.03 i 0.36 ± 0.04, respectivament. Tot i això, aquesta millora no és significativa, i ve acompanyada d'un cost important: la pèrdua de moltes mostres de vídeo. Tal com es va evidenciar a la Fig.7, reduir el nombre de vídeos pot limitar greument la capacitat del model per capturar la variabilitat intra-classe i generalitzar de manera efectiva. Per tant, encara que el downsampling redueix el risc de sobreajustament, no resulta òptim en escenaris amb escassetat de dades, on el volum de mostres és clau per assolir un rendiment robust.

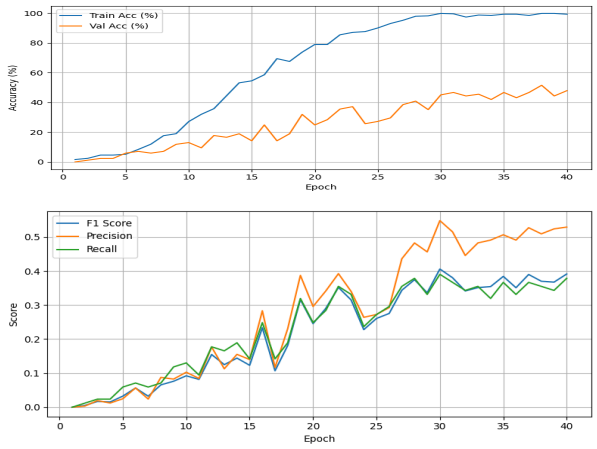


Fig 11. Corbes d'accuracy, precision, recall i F1 score obtingudes amb l'estratègia downsampling durant l'entrenament.

4.8 Resultats

El model definitiu es basa en una arquitectura BLSTM amb atenció MultiHead, que ha demostrat ser la combinació més efectiva, especialment pel que fa al recall (0.42 ± 0.03). L'ús de Dropout i LR adaptatius ha contribuït a reduir l'overfitting i a afavorir la convergència. Tot i que l'estratègia de DownSampling ha reduït el desbalanceig, no dona evidencia estadística de millora. El finetuning parcial de la ResNet18 ha millorat la qualitat de les representacions visuals amb un cost computacional reduït. Pel que fa al batch size, la seva configuració reduïda (batch de 4) ha estat una limitació significativa, ja que ha introduït més soroll en l'optimització i n'ha dificultat l'avaluació sistemàtica. En conjunt, aquestes decisions han permès obtenir el model més robust dels experimentats, tot i identificar àrees clau de millora per a futurs treballs.

5 CONCLUSIONS

Aquest treball ha assolit els objectius inicials, desenvolupant un sistema funcional capaç de reconèixer signes en vídeos mitjançant tècniques de visió per computador i aprenentatge profund. Concretament, s'ha implementat una arquitectura CNN+BLSTM amb mecanismes d'atenció i regularització, que ha permès capturar de manera eficient la dinàmica temporal de les seqüències. El sistema ha estat capaç de traduir vídeos a glosses escrites amb una accuracy de validació de fins a 40 % ± 2.4 %, una precisió mitjana (Precision) de 0.44 ± 0.06, un recall de 0.42 ± 0.03 i un F1 Score de 0.36 ± 0.04.

A nivell de processament multimodal, s'ha integrat un sistema de detecció de mans i cares basat en YOLOv8 amb una precisió de 99.66 % i un recall de 99.52 %, cosa que ha permès centrar l'atenció del model en les regions més rellevants dels vídeos. Aquest enfocament, combinat amb tècniques de normalització i augmentació espaciotemporal, ha millorat la qualitat de les dades d'entrada.

Pel que fa a l'anàlisi contextual i temporal, s'han emprat xarxes BLSTM bidireccionals, pooling híbrid i codificació posicional, així com mòduls d'atenció. Els resultats mostren que l'atenció Multi-Head ha estat més efectiva, especialment en el recall, millorant la detecció de classes minoritàries.

Tot i aquestes millores, els resultats també han posat de manifest diverses limitacions. D'una banda, el conjunt de dades WLASL presenta un fort desequilibri entre classes i molt poques mostres per vídeo, dificultant l'aprenentatge generalitzat. Tècniques com l'oversampling han degradat el rendiment per excés de redundància (amb val. accuracy ~27 %), mentre que el downsampling, tot i millorar lleugerament el rendiment, ha limitat la variabilitat i el potencial de generalització. De l'altra, la necessitat computacional per entrenar aquest tipus de models és molt elevada. L'ús de batch size petits (4), per restriccions de memòria, ha introduït soroll en l'optimització i dificultat l'aprenentatge de patrons globals.

Malgrat tot, aquest projecte estableix una base sòlida per seguir avançant en el reconeixement automàtic de llengua de signes. Les tècniques implementades, els resultats obtinguts i l'anàlisi crítica de les estratègies emprades aporten valor tant a nivell tècnic com metodològic, obrint la porta a millores futures mitjançant l'escalat del corpus, la integració de models lingüístics i una optimització més eficient. Per tant, es pot concloure que els objectius plantejats han estat assolits en gran part, demostrant que, fins i tot amb recursos limitats i dades imperfectes, es poden obtenir resultats prometedors en aquest àmbit.

6 TREBALL FUTUR

Per finalitzar aquest treball, m'agradaria mencionar tres propostes que poden servir com a línies de continuació per aprofundir i ampliar aquesta investigació:

6.1 Integració Social:

Aquest sistema pot ser una eina clau per millorar la comunicació de les persones amb discapacitats auditives, promovent una societat més accessible i justa. A més, pot impulsar iniciatives tecnològiques orientades a col·lectius vulnerables, fomentant la igualtat d'oportunitats.

6.2 Rames d'investigació:

De cara al futur, cal explorar l'entrenament del model amb més capacitat computacional per analitzar l'impacte

del batch size i la profunditat del model en el rendiment. També és interessant integrar tècniques de Natural Language Processing (NLP) per generar frases coherents a partir dels signes detectats, avançant cap a sistemes complets de traducció automàtica.

6.3 Base de dades Catalana:

Un pas interessant seria la creació o adaptació d'un corpus específic per a la Llengua de Signes Catalana (LSC), ja que actualment hi ha una clara mancança de recursos públics i anotats per a aquesta llengua. La disponibilitat d'un conjunt de dades en LSC permetria desenvolupar models adaptats al context sociolingüístic local i impulsar tecnologies inclusives per a la comunitat sorda catalana. Això obriria la porta a projectes d'alt impacte social i cultural, promovent la preservació i digitalització de la LSC.

7 Agraïments

En finalitzar, voldria expressar el meu més sincer agraïment a totes aquelles persones que m'han acompanyat i donat suport al llarg d'aquest procés.

En primer lloc, a la Gemma Sánchez, per fer possible la realització d'un projecte tan enriquidor i amb una finalitat tan significativa. La seva orientació, disponibilitat i implicació han estat claus per al desenvolupament d'aquest treball.

També vull agrair especialment a l'Eduard Ara a l'Àlex Guareño, dos referents fonamentals al llarg del meu pas per la universitat, per la seva constància de superació, el seu suport i la seva amistat.

Finalment, a la meva família, per ser el pilar que sempre m'ha sostingut. Per haver-me transmès els valors, l'esforç i la perseverança que m'han ajudat a arribar fins aquí. Moltes gràcies a tots; sense vosaltres, aquest projecte no hauria estat possible.

8 Bibliografia

[1] World Federation of the Deaf, About us, [En línia]. Disponible a: https://wfdeaf.org

[2] CNSE, Confederación Estatal de Personas Sordas, [En línia]. Disponible a: https://www.cnse.es

[3] Brentari, D. (Ed.). Sign Languages. Cambridge University Press, 2010.

[4] Koller, O. (2020). Quantitative survey of the state of the art in sign language recognition. arXiv preprint arXiv:2008.09918v2. https://arxiv.org/abs/2008.09918

[5]]Baskoro, R. (n.d.). WLASL-Processed Dataset. Kaggle. Recuperado de https://www.kaggle.com/datasets/risangbaskoro/WLASL-Processed

[6] Li, D. (n.d.). WLASL: A large-scale dataset for Word-Level American Sign Language. GitHub. Recuperado de https://github.com/dxli94/WLASL

[7] Cavalcante Carneiro, A. L. (2023). Hand and face detection focused on sign language [Dataset]. Kaggle. https://www.kaggle.com/datasets/alvarole/hand-and-face-detection-focused-on-sign-language

[8] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009).ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE.https://doi.org/10.1109/CVPR.2009.5206848

[9] Simonyan, K., & Zisserman, A. (2015).Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR).

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016).Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[11] YOLOv8 Nano vs. YOLOv8 Small – Roboflow https://roboflow.com/compare-model-sizes/yolov8-nano-vs-yolov8-small

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[13] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.0473

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1706.03762

[15] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1506.07503

[16] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

[18] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," Neural Networks, vol. 113, pp. 54–71, 2019. [Online]. Available: https://doi.org/10.1016/j.neunet.2019.01.012

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf

[20] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural Networks, vol. 106, pp. 249–259, 2018. [Online]. Available: https://doi.org/10.1016/j.neunet.2018.07.011

[21] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural Networks, vol. 106, pp. 249–259, 2018. [Online]. Available: https://doi.org/10.1016/j.neunet.2018.07.011

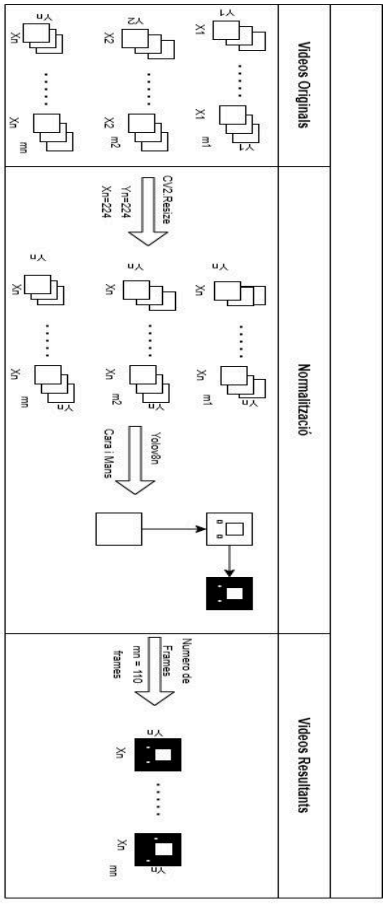
[22] L. Smith, "Cyclical learning rates for training neural networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2017, pp. 464–472. [Online]. Available: https://arxiv.org/abs/1506.01186

[23] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

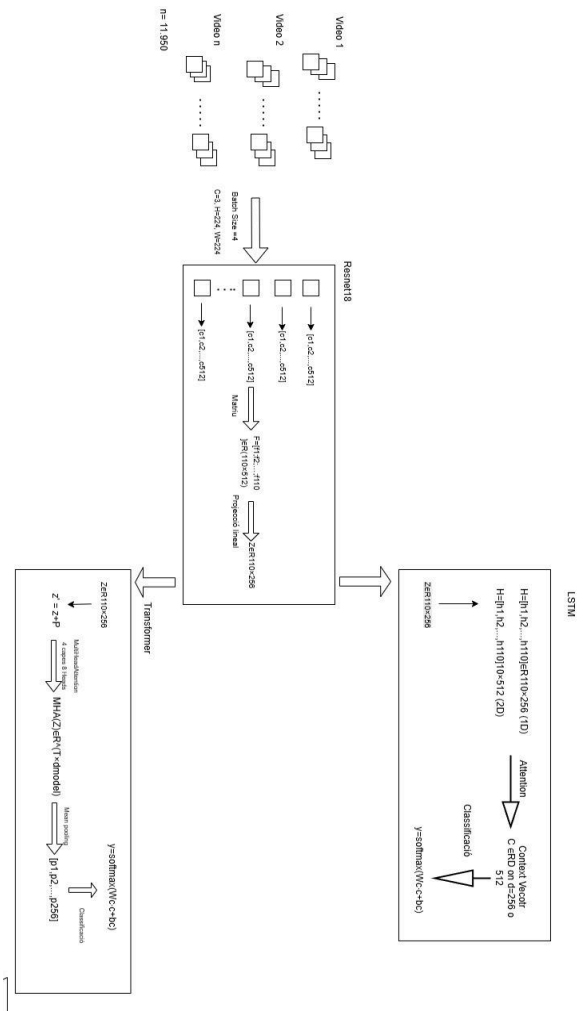
[24] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition," arXiv preprint arXiv:2002.03187, 2020. [Online]. Available: https://arxiv.org/abs/2002.03187

APÈNDIX

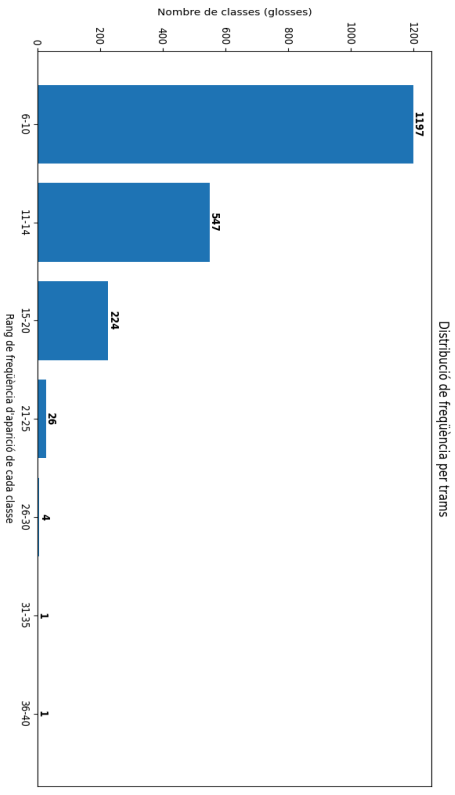
A.1 Diagrama del tractament de dades



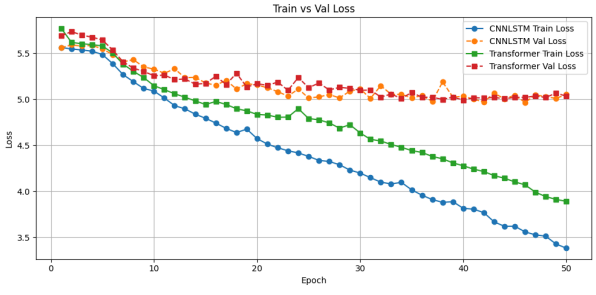
A.2 Diagrama del model de Deep Learning



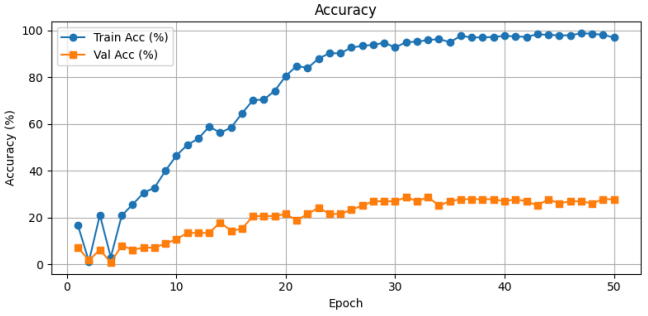
A.3 Distribució del nombre de vídeos per classe en el conjunt de dades.



A.4 Loss del model provat a la (fig 2) (Transformers VS LSTM 4.1)



A.5 Accuracy del model provat al aprtat (Millores BLSTM 4.3)



A.6 Distribució de quantitat de video mitjançant reequilibri de les dades.

Min Frecuencia	Original	Downsampling	Oversampling
15 (256 clases)	4501	3840	10240
20 (50 clases)	1131	1020	2040
25(8 clases)	233	200	320