Deepfake Detection Through Facial Dynamics

Fondamenti di Visione Artificiale e Biometria - GRUPPO 4

Danilo Gisolfi
Università degli Studi di Salerno
Dipartimento di Informatica
0522502001

Vincenzo Maiellaro
Università degli Studi di Salerno
Dipartimento di Informatica
0522502055

Tommaso Nardi
Università degli Studi di Salerno
Dipartimento di Informatica
0522502035

Abstract

Negli ultimi anni, la diffusione dei deepfake — video sintetici ottenuti tramite tecniche avanzate di intelligenza artificiale — ha sollevato importanti problematiche di sicurezza, disinformazione e violazione della privacy.

Il presente studio propone un sistema innovativo per il rilevamento automatico di video deepfake, basato sull'analisi delle dinamiche facciali con particolare attenzione ai movimenti labiali. L'architettura sviluppata integra reti neurali convoluzionali (CNN) per l'estrazione di caratteristiche spaziali e reti Long Short-Term Memory (LSTM) per la modellazione delle dipendenze temporali, creando un approccio ibrido capace di identificare incoerenze nei contenuti video.

L'approccio metodologico ha seguito una strategia incrementale di ottimizzazione: partendo da una configurazione iniziale con analisi dell'intero volto, si è evoluto verso una soluzione ottimizzata focalizzata esclusivamente sulla regione labiale. L'innovazione principale risiede nell'implementazione di una tecnica di selezione dei frame basata sull'Optical Flow, che permette di identificare e analizzare i momenti di maggiore attività dinamica migliorando l'efficacia del rilevamento.

Il sistema sviluppato apre prospettive per future estensioni verso il rilevamento multimodale e l'implementazione in contesti real-time, contribuendo alla lotta contro la disinformazione digitale e supportare una comunicazione più trasparente e verificabile.

1 INTRODUZIONE

1.1 Contesto e motivazioni

L'avanzamento esponenziale delle tecnologie di intelligenza artificiale ha dato vita a una nuova era nel mondo della manipolazione mediatica: i *deepfake*. Questi video sintetici, generati attraverso sofisticate architetture di reti neurali sono diventati così realistici da rendere estremamente difficile distinguere un contenuto autentico da uno manipolato. Il fenomeno non è più limitato alle sperimentazioni come quelle in campo accademico, ma si sta ormai diffondendo su scala globale, con implicazioni preoccupanti in diversi settori.

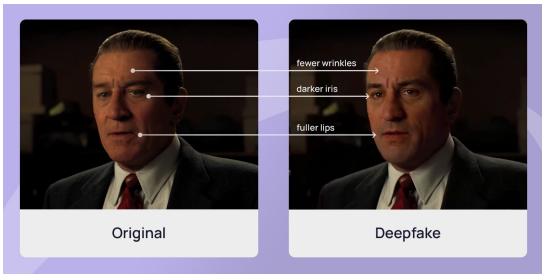

Figure 1: Confronto tra un volto originale (a sinistra) e la sua versione deepfake (a destra). ***Fonte: Shamook (2020), De-aging Robert De Niro in The Irishman [DeepFake].***

Come illustrato in Figura 1, la capacità di alterare volti e mimica facciale con una fedeltà quasi perfetta pone sfide senza precedenti. La manipolazione di espressioni, allineamento labiale e micromovimenti facciali è diventata così convincente da rendere difficile per l'occhio umano percepire le incoerenze. Questa rapida evoluzione tecnologica solleva questioni critiche e minacce significative su più fronti:

- *Sicurezza e disinformazione*: La diffusione di video falsificati può essere utilizzata per minare la fiducia in figure pubbliche e istituzioni democratiche, manipolando l'opinione pubblica attraverso la creazione di dichiarazioni mai pronunciate o eventi fittizi.
- *Cybersicurezza*: I deepfake possono rappresentare una seria minaccia ai sistemi di autenticazione biometrica basati sul riconoscimento facciale o vocale, facilitando attacchi di falsificazione di identità.
- *Violazione della Privacy e Cyberbullismo*: La creazione di contenuti falsi e compromettenti può essere impiegata per danneggiare la reputazione di individui, mettere in atto estorsioni o alimentare campagne di cyberbullismo, con gravi ripercussioni personali e sociali.

1.2 Obiettivi e approccio

Di fronte a queste problematiche, si rende indispensabile lo sviluppo di sistemi automatici robusti e affidabili per la rilevazione dei deepfake. Il presente lavoro si propone di affrontare questa sfida critica attraverso un approccio basato sull'analisi delle dinamiche facciali. Il sistema proposto si fonda su una potente architettura ibrida che combina Convolutional Neural Networks (CNN) e Long Short-Term Memory (LSTM). Le CNN sono impiegate per estrarre

[1] *De-aging Robert De Niro in The Irishman [DeepFake]*, https://www.youtube.com/watch?v=dHSTWepkp_M&ab_channel=Shamook

caratteristiche spaziali dettagliate da singoli frame video, mentre le LSTM sono cruciali per modellare le dipendenze temporali e le sequenze di movimento nel tempo. Questa sinergia permette di analizzare non solo l'aspetto statico, ma anche le sottili incoerenze dinamiche tipiche dei video generati artificialmente. L'analisi si concentrerà specificamente sulle regioni facciali e, in particolare, sui movimenti labiali. Questa scelta è motivata dal fatto che i deepfake, pur avendo raggiunto un elevato livello di realismo, spesso presentano artefatti o incongruenze meno evidenti, ma persistenti proprio nelle dinamiche labiali, quali la sincronizzazione con l'audio o la naturalezza dei movimenti. L'obiettivo finale di questo progetto è sviluppare un sistema di rilevamento dei deepfake che sia non solo accurato ed efficace, ma anche robusto e generalizzabile, capace di operare in contesti reali e su un'ampia varietà di contenuti manipolati. Contribuendo a migliorare la capacità di discernere tra realtà e manipolazione digitale, intendiamo rafforzare la lotta contro la disinformazione e tutelare l'integrità del panorama mediatico nell'era digitale.

2 RELATED WORKS

Tale capitolo presenta una panoramica generale delle attuali tecnologie e metodologie utilizzate per il rilevamento di video deepfake, analizzando sia gli approcci basati su caratteristiche biometriche sia quelli fondati sull'analisi dei pattern di codifica video.

A conferma dell'importanza dell'integrazione tra dimensione spaziale e temporale, molti degli approcci più recenti si fondano su architetture ibride CNN-LSTM. Queste combinano la capacità delle CNN di estrarre rappresentazioni visive dettagliate dai singoli frame con quella delle LSTM di modellare la dinamica temporale all'interno delle sequenze.

Le tecniche di rilevamento dei deepfake più efficaci sviluppate negli ultimi anni si basano principalmente su architetture di deep learning, in particolare sulle reti neurali convoluzionali (CNN) e sulle architetture ibride che le integrano con moduli temporali. Le CNN si distinguono per la loro abilità di apprendere automaticamente rappresentazioni gerarchiche dei dati visivi, analizzando le immagini attraverso diversi livelli di astrazione. Nei primi strati convoluzionali il modello rileva caratteristiche di basso livello come bordi, angoli e texture locali; nei livelli intermedi questi pattern vengono combinati per riconoscere strutture più complesse, come porzioni di volti (occhi, bocca, naso); negli strati più profondi le CNN apprendono rappresentazioni ad alto livello come configurazioni facciali complete, espressioni o movimenti anomali. Questa stratificazione progressiva dell'informazione visiva consente a queste reti neurali di cogliere dettagli impercettibili all'occhio umano e di identificare in modo efficace le sottili incoerenze o artefatti tipici dei contenuti manipolati.

Invece, le Long Short-Term Memory (LSTM) sono una variante delle reti neurali ricorrenti (RNN) progettata per superare le difficoltà delle RNN tradizionali nel mantenere informazioni su lunghe sequenze. Grazie alla loro struttura interna (celle di memoria, porte di ingresso, uscita e dimenticanza) sono in grado di apprendere dipendenze temporali a lungo termine, trattenendo o scartando selettivamente informazioni nel tempo. In ambito video, questo si traduce nella possibilità di modellare l'evoluzione temporale dei movimenti facciali o delle transizioni tra espressioni, elementi cruciali per identificare manipolazioni sottili.

Un esempio rappresentativo di questa combinazione è CLRNet (Convolutional LSTM Residual Network), introdotto da Tariq et al., che unisce una CNN residuale con moduli LSTM per analizzare sequenze brevi di fotogrammi. CLRNet è stato progettato per catturare artefatti inter-frame, come incoerenze nell'illuminazione e micro-scatti facciali, ed è stato valutato sul dataset FaceForensics++, ottenendo un'accuratezza del 97,57% e dimostrando una buona generalizzazione anche su video deepfake non visti precedentemente [1]

Mitra et al. hanno proposto un'architettura end-to-end completamente connessa che utilizza XceptionNet CNN come estrattore di caratteristiche dai fotogrammi video. Il loro approccio si basa sul principio che se anche un solo fotogramma viene classificato come contraffatto, l'intero video viene considerato fake. L'addestramento e il test sono stati eseguiti sul dataset Celeb-DF, noto per la sua complessità e realismo. I risultati mostrano ottimi risultati, con un'accuratezza massima del 98,26% [2].

Un altro contributo significativo proviene da Awotunde et al., che hanno sviluppato un modello di rilevamento e classificazione deepfake basato su CNN a cinque strati. Il loro approccio, potenziato con funzioni di attivazione ReLU, estrae caratteristiche dalle regioni facciali nei fotogrammi video mentre l'addestramento e la valutazione sono effettuati sul dataset Deepfake Detection Challenge Dataset [3].

Un filone particolarmente promettente nella ricerca sul rilevamento dei deepfake si basa sull'analisi dei tratti biometrici univoci dell'individuo, sfruttando la coerenza tra le caratteristiche comportamentali e fisiologiche di un soggetto. In questo contesto si inserisce POIForensics, un framework sviluppato da un gruppo di ricercatori italiani e tedeschi, che utilizza come segnali di riferimento i movimenti facciali e le componenti audio di un individuo reale sottoposto a manipolazione. A differenza di molti metodi che richiedono un addestramento specifico per ogni soggetto, POIForensics necessita solamente di una decina di video autentici per costruire un profilo di riferimento, sfruttando distanze vettoriali già apprese dal modello. Questo lo rende particolarmente flessibile e capace di generalizzare anche in presenza di attacchi non visti. Inoltre, l'approccio è in grado di gestire sia manipolazioni unimodali (solo audio o solo video), sia attacchi multimodali, rivelandosi efficace in scenari complessi e realistici [4].

Oltre agli approcci basati sui tratti biometrici, altri studi si sono concentrati sull'analisi delle caratteristiche intrinseche della codifica video, soprattutto in contesti in cui i dati disponibili per l'addestramento sono limitati. Un esempio rilevante è il metodo proposto da Wang et al., che analizza le modalità di predizione del movimento nei video sfruttando una sorta di "impronta digitale" codificata. La loro osservazione chiave è che i video deepfake presentano schemi di movimento diversi rispetto ai video autentici, risultando spesso meno prevedibili e con una maggiore incidenza di frame codificati in modalità intra, ovvero frame che vengono compressi senza fare riferimento ad altri fotogrammi. Basandosi su queste informazioni, il sistema utilizza un classificatore SVM lineare che, pur facendo uso di caratteristiche semplici, riesce a ottenere un'elevata accuratezza anche con dataset ridotti [5].

3 ARCHITETTURA DEL SISTEMA PROPOSTO E PIPELINE DI ELABORAZIONE DATI

3.1 Introduzione e panoramica generale

Il presente capitolo illustra in dettaglio l'architettura e la metodologia del sistema proposto per la rilevazione automatica di video deepfake. Come anticipato nel Capitolo introduttivo, l'obiettivo principale è sviluppare una soluzione efficace per distinguere video autentici da quelli manipolati con l'intelligenza artificiale, concentrandosi soprattutto sui movimenti delle labbra, area spesso soggetta a errori nei video sintetici.

Il progetto ha seguito un approccio incrementale, partendo da configurazioni di base per l'estrazione dei frame e delle regioni di interesse (ROI) fino all'implementazione di tecniche più sofisticate. Il sistema finale si basa su un'architettura ibrida che integra reti neurali convoluzionali (CNN) per l'estrazione di feature spaziali e reti Long Short-Term Memory (LSTM) per la modellazione delle dipendenze temporali, approccio che combina l'analisi delle immagini e dei movimenti per rilevare anomalie nei deepfake.

Il flusso di lavoro complessivo del sistema, dalle fasi di preprocessing all'inferenza, si articola come segue:

(1) *Pre-processing dei dati:* I video originali vengono elaborati per estrarre sequenze di frame. Questa fase include diverse strategie: dall'estrazione uniforme dei frame, al rilevamento e ritaglio delle regioni di interesse (volti e labbra) tramite MediaPipe, fino alla selezione avanzata dei frame basata sull'analisi dell'Optical Flow.
(2) *Estrazione delle feature spaziali (CNN):* Le sequenze di frame (già ritagliate sulla regione delle labbra e ridimensionate) vengono passate attraverso una componente convoluzionale (CNN) che apprende automaticamente rappresentazioni gerarchiche delle caratteristiche visive.
(3) *Modellazione temporale (LSTM):* Le feature spaziali estratte vengono inserite in una rete LSTM, che è in grado di catturare le dipendenze a lungo termine e le dinamiche temporali all'interno delle sequenze di frame.
(4) *Classificazione:* L'output della rete LSTM viene infine elaborato da un modulo di classificazione che determina se il video è autentico o manipolato.

Questa architettura modulare, migliorata attraverso diverse fasi di ottimizzazione, è stata progettata per bilanciare accuratezza predittiva e efficienza computazionale, rendendola adatta all'identificazione di alterazioni sottili e costanti nel tempo.

3.2 Dati e materiali utilizzati

Per lo sviluppo, l'addestramento e la valutazione del sistema sono stati impiegati diversi dataset di video, oltre a un ambiente di sviluppo e librerie specifiche.

3.2.1 Dataset di addestramento e valutazione

La selezione dei dataset è stata cruciale per garantire la robustezza e la capacità di generalizzazione del modello, esplorando diverse configurazioni di input.

- *Celeb-DFv2*[1] *:* Questo dataset è una delle risorse pubbliche più complete per lo studio dei deepfake. Il dataset è composto da 590 video originali, con soggetti appartenenti a diversi gruppi etnici, fasce di età e generi, e da 5639 video deepfake generati tramite tecniche avanzate di sostituzione facciale. Le manipolazioni presenti nei video sono di qualità particolarmente elevata, rendendo il compito di classificazione piuttosto difficile e quindi ideale per la valutazione di modelli robusti.
- *M2FRED*[2] */ WAV2LIP*[3] *:* Questi sono due dataset complementari: *M2FRED* contiene video reali e non manipolati, mentre *WAV2LIP* è costituito da video deepfake. L'utilizzo congiunto di questi due dataset permette di valutare la capacità del sistema di distinguere tra contenuti video autentici e quelli artificialmente alterati.
- *XM2VTS*[4]: In questo dataset tutti i video sono reali e non contengono manipolazioni. Per tale ragione, il suo impiego è stato mirato a verificare se i modelli addestrati sui dataset precedenti (Celeb-DFv2, M2FRED / WAV2LIP) fossero in grado di identificare correttamente i video di XM2VTS come autentici, valutando così la loro performance su dati reali non visti durante l'addestramento.

La tabella seguente riassume la composizione dei dataset utilizzati:

Table 1: Panoramica dei Dataset di Video

Dataset	*Video real*	*Video fake*
Celeb DFv2	590	5639
M2Fred/WAV2LIP	611	2635
XM2VTS	1208	0

3.2.2 Ambiente di sviluppo e librerie

L'implementazione del sistema è stata realizzata utilizzando diverse librerie e un ambiente di sviluppo cloud.

- *OpenCV:* Libreria open-source per la visione artificiale e l'elaborazione delle immagini. È stata impiegata ampiamente nelle fasi di pre-processing per la lettura e scrittura dei video/frame, la conversione tra spazi colore, il calcolo dell'Optical Flow e le operazioni di ritaglio/ridimensionamento delle immagini.
- *MediaPipe:* Un framework di machine learning multimodale sviluppato da Google. È stato utilizzato per il rilevamento e il tracking accurato dei volti e dei landmark facciali nei video. Questo è cruciale per l'estrazione delle Regioni di Interesse (ROI) di volto e labbra, consentendo un'identificazione robusta anche in condizioni variabili di illuminazione e posa. MediaPipe è un componente chiave nella fase di estrazione delle ROI.

[1] CELEB: https://www.kaggle.com/datasets/jordangray/celeb-df-v2
[2] M2FRED: https://github.com/ondyari/FaceForensics
[3] WAV2LIP:https://github.com/Rudrabha/Wav2Lip
[4] XM2VTS: https://www.idiap.ch/webarchives/sites/www.idiap.ch/resource/faceverif/

- *TensorFlow / Keras:* TensorFlow è il framework open-source per il machine learning e il deep learning di Google, utilizzato come backend principale per l'implementazione e l'addestramento dell'architettura CNN-LSTM. Keras, come API di alto livello, ha semplificato la costruzione e l'addestramento dei modelli di deep learning.
- *Ambiente di esecuzione:* Il sistema è stato sviluppato ed eseguito sulla piattaforma cloud *Kaggle*, utilizzando una macchina virtuale con supporto a Jupyter Notebook. Le specifiche tecniche tipiche delle macchine virtuali utilizzate includevano:
  - *GPU:* 2 × NVIDIA Tesla T4, ciascuna con 16 GB di VRAM GDDR6 (per un totale 32 GB).
  - *CPU:* 2 × Intel Xeon @ 2.2 GHz.
  - *RAM:* 32 GB.

  La scelta di Kaggle ha permesso di accedere a risorse computazionali significative e a librerie preinstallate, ottimizzando il processo di sviluppo e sperimentazione.

3.3 Pipeline di Pre-processing dei dati

Il pre-processing dei dati è una fase critica per preparare i video in un formato adatto all'input del modello. Nel corso del progetto, sono state esplorate e implementate diverse strategie di estrazione dei frame e delle Regioni di Interesse (ROI), adottando un approccio graduale per migliorarne l'efficacia.

3.3.1 Estrazione uniforme dei frame

In una prima configurazione, i frame sono stati estratti dai video in modo uniforme. Questo processo è gestito da un modulo dedicato che apre il video, calcola il numero totale di frame e determina uno "step" per selezionare uniformemente 60 frame (o un numero predefinito) lungo l'intera durata del video. I frame selezionati vengono salvati in cartelle di output dedicate.

3.3.2 Estrazione delle ROI di volto e labbra

Successivamente all'estrazione dei frame, è stata implementata una fase dedicata al rilevamento e al ritaglio delle Regioni di Interesse (ROI) più rilevanti: il volto e, in particolare, le labbra. Questa fase è gestita da un modulo specifico che utilizza *MediaPipe*:

- Per ogni frame estratto, MediaPipe viene impiegato per rilevare il volto e i relativi landmark facciali.
- Una volta identificato il volto, la sua bounding box viene utilizzata per ritagliare la ROI del volto.
- A partire dalla ROI del volto e utilizzando specifici landmark facciali (es. quelli che definiscono il contorno delle labbra, come 61, 291, 0, 17, 13, 14, 78, 308), viene calcolata e ritagliata la ROI delle labbra.

Questa fase consente poi al modello di analizzare le zone del viso più vulnerabili ai deepfake.

3.3.3 Selezione frame basata su Optical Flow

Per migliorare ulteriormente l'efficacia del pre-processing, è stata introdotta una tecnica di selezione dinamica dei frame basata sull'*Optical Flow*. L'obiettivo è selezionare i frame che presentano la maggiore attività di movimento, in quanto sono più propensi a contenere informazioni critiche per rilevare anomalie.

- *Concetto di Optical Flow:* L'*Optical Flow* è una tecnica che si occupa di stimare il movimento apparente degli oggetti tra fotogrammi consecutivi in una sequenza video [6]. In termini pratici, produce un campo vettoriale che descrive lo spostamento di ogni singolo pixel da un frame al successivo. Il principio fondamentale su cui si basa questa tecnica è la *conservazione della luminosità*, secondo cui si assume che l'intensità luminosa di un punto dell'immagine rimanga pressoché costante nel tempo. Tale assunzione può essere formalizzata mediante la seguente relazione:

  $$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$

  dove $I(x, y, t)$ rappresenta l'intensità luminosa del pixel alle coordinate (x, y) al tempo t, mentre Δx e Δy indicano il suo spostamento nei successivi intervalli temporali.
  Sviluppando tale relazione mediante un'espansione in serie di Taylor e trascurando i termini di ordine superiore, si ottiene l'equazione del vincolo di Optical Flow:

  $$I_x \cdot u + I_y \cdot v + I_t = 0$$

  dove:
  - I_x e I_y sono le derivate spaziali dell'immagine rispetto a x e y,
  - I_t è la derivata temporale,
  - u e v sono le componenti del vettore di spostamento (flusso ottico) nelle direzioni orizzontale e verticale.
- *Calcolo e Misura dell'Intensità:* Il calcolo del campo di Optical Flow viene effettuato tra coppie di frame consecutivi (convertiti in scala di grigi) utilizzando l'algoritmo di *Farneback*, come implementato nella libreria OpenCV. Per ogni pixel, si ottiene un vettore di spostamento, da cui si calcola la *magnitudine del movimento* tramite la formula:

  $$M(x, y) = \sqrt{u(x, y)^2 + v(x, y)^2}$$

  L'intensità complessiva del movimento per una coppia di frame viene poi stimata calcolando la media della magnitudine su tutti i pixel:

  $$\bar{M} = \frac{1}{N} \sum_{x,y} M(x, y)$$

  dove N è il numero totale di pixel nel frame.
- *Criterio di Selezione:* Vengono selezionati solo i frame la cui intensità media del movimento $\bar{M}$ supera una soglia predefinita $\theta = 1.0$, valore determinato empiricamente. I frame che soddisfano questo criterio vengono selezionati in base all'intensità, scegliendo i primi 60 con maggiore attività.

L'utilizzo dell'Optical Flow si è rivelato particolarmente efficace, poiché i deepfake spesso presentano anomalie nei micro-movimenti facciali o nella sincronizzazione labiale che risultano più evidenti nei momenti di maggiore attività. Concentrando l'analisi su questi frame, si riduce il rumore computazionale e si migliora l'efficienza di apprendimento del modello.

3.4 Architettura del Modello CNN-LSTM

Il cuore del sistema proposto è un'architettura di deep learning ibrida CNN-LSTM, progettata per sfruttare al meglio sia le caratteristiche spaziali delle immagini che le dipendenze temporali nelle sequenze video.

3.4.1 Panoramica generale

L'architettura riceve in input sequenze di *60 fotogrammi* per video, ciascuno rappresentante la regione delle labbra con una risoluzione di *64x64 pixel e 3 canali RGB*. L'obiettivo è produrre una classificazione binaria: determinare se il video è "autentico" o "deepfake".

3.4.2 Componente Convoluzionale (CNN)

La componente CNN è responsabile dell'estrazione automatica delle feature spaziali significative da ciascun frame della sequenza. Per consentire alla CNN di elaborare ogni frame in modo indipendente ma all'interno del contesto sequenziale, viene utilizzato il wrapper *TimeDistributed* di Keras, che applica lo stesso set di pesi convoluzionali a tutti i frame della sequenza.

La CNN è composta da due blocchi convoluzionali principali:

- *Primo Blocco:*
  - Un layer *Conv2D* con *32 filtri*, un kernel di dimensione *3x3* e funzione di attivazione *ReLU*. Questo strato apprende feature di basso livello come bordi e texture.
  - Seguito da un layer *MaxPooling2D* con un kernel *2x2*, che riduce la dimensionalità spaziale delle feature map e contribuisce a rendere il modello più robusto a piccole variazioni di posizione.
- *Secondo Blocco:*
  - Un layer *Conv2D* con *64 filtri*, un kernel *3x3* e attivazione *ReLU*. Questo strato apprende feature di livello superiore, combinando i pattern rilevati dal primo blocco. È applicata una regolarizzazione L2 sui pesi.
  - Seguito anch'esso da un layer *MaxPooling2D* con un kernel *2x2*, riducendo ulteriormente la dimensionalità.

Le feature map prodotte da questi blocchi vengono infine appiattite tramite un layer *Flatten*, che le trasforma in vettori monodimensionali. Questi vettori di feature per ciascun frame vengono poi passati alla componente LSTM.

3.4.3 Componente Temporale (LSTM)

La componente LSTM (Long Short-Term Memory) è essenziale per modellare le dipendenze temporali tra i frame e per identificare anomalie che emergono nel corso del video. Le LSTM sono particolarmente adatte a gestire sequenze di dati, grazie alle loro "celle di memoria" e ai "gate" (input, forget, output) che permettono di mantenere o scartare informazioni a lungo termine.

- *Struttura:* È implementato un singolo layer *LSTM* con *128 unità*.
- *Dropout:* Vengono applicati dropout sia sull'input che sullo stato ricorrente per migliorare la regolarizzazione, una tecnica che aiuta a prevenire l'overfitting disattivando casualmente alcune connessioni durante l'addestramento.
- *Output:* Il layer LSTM è configurato per restituire solo l'output dell'ultimo timestep della sequenza. Questo output, un vettore di 128 dimensioni, rappresenta una codifica compressa e riassuntiva dell'intera sequenza temporale del video, catturando le dinamiche e le relazioni tra i frame.

La capacità della LSTM di apprendere dipendenze a lungo termine è cruciale per la rilevazione dei deepfake, poiché incoerenze nei movimenti labiali, nelle espressioni o nella sincronizzazione possono manifestarsi solo osservando il video nella sua interezza temporale e non solo su frame isolati.

3.4.4 Modulo di Classificazione

L'output del layer LSTM (il vettore di 128 dimensioni) viene passato al modulo di classificazione, che ha il compito di produrre la previsione finale (autentico o deepfake).

- *Primo Layer Dense:* Un layer completamente connesso con *128 neuroni* e funzione di attivazione *ReLU*. È applicata una regolarizzazione L2 sui pesi.
- *Layer Dropout:* Un layer di *Dropout* viene applicato con una percentuale del *50%*.
- *Layer di Output:* Un layer *Dense* finale con *1 neurone* e funzione di attivazione *Sigmoid*. Questo layer produce un valore tra 0 e 1, che può essere interpretato come la probabilità che il video sia un deepfake (es. valori vicini a 1 indicano deepfake, valori vicini a 0 indicano autentico).

3.5 Processo di addestramento e ottimizzazione

Il processo di addestramento del modello è stato attentamente ottimizzato per garantire stabilità, efficienza e robustezza, minimizzando l'overfitting e massimizzando la capacità di generalizzazione.

- *Funzione di Loss:* È stata utilizzata la *Binary Crossentropy*, una funzione di loss standard per problemi di classificazione binaria, che misura la divergenza tra le probabilità predette e le etichette reali.
- *Metriche di Valutazione:* Durante l'addestramento e la valutazione, sono state monitorate diverse metriche chiave: *Accuracy, Precision, Recall*.
- *Tecniche di Regolarizzazione:*
  - *Dropout:* Implementato sia nel layer LSTM (input e recurrent dropout) che nel modulo di classificazione, per impedire al modello di fare eccessivo affidamento su specifici percorsi e migliorare la sua capacità di generalizzazione.
  - *L2 Regularization:* Applicata ai layer convoluzionali e Dense, penalizza i pesi di grandi dimensioni, scoraggiando l'overfitting.
- *Callbacks:* È stato integrato un callback per ottimizzare il processo di training:
  - *Early Stopping:* Monitora la loss di validazione e interrompe automaticamente l'addestramento se questa non migliora per 5 epoche, ripristinando i pesi migliori. Questo previene l'overfitting e risparmia tempo computazionale.
- *Ottimizzazioni per l'Efficienza Computazionale:*
  - *Mixed Precision:* È stata adottata la tecnica di mixed precision, che utilizza numeri a virgola mobile a 16 bit per i calcoli intermedi e a 32 bit per i layer di output. Questo riduce significativamente l'utilizzo della memoria GPU

e accelera l'addestramento, mantenendo al contempo la stabilità numerica necessaria.
  - *tf.data.Dataset:* Per gestire grandi volumi di dati e ottimizzare il caricamento, è stato utilizzato un modulo di TensorFlow per la creazione di dataset. Questo ha permesso un caricamento progressivo dei dati senza la necessità di caricare l'intero dataset in RAM. Il modulo supporta funzionalità come caching (per memorizzare i dati preprocessati), shuffling (per randomizzare i dati) e prefetching (per sovrapporre il pre-processing all'esecuzione del modello), ottimizzando l'efficienza del training. Il batch size è impostato a 4.
- *Data Augmentation:* Per aumentare la robustezza e la capacità di generalizzazione del modello, sono state applicate tecniche di data augmentation ai frame di input durante il pre-processing del dataset, tra cui:
  - *Random Horizontal Flip:* Capovolgimento orizzontale casuale delle immagini.
  - *Random Brightness Adjustment:* Variazioni casuali della luminosità entro un intervallo predefinito.

Questa combinazione di tecniche di addestramento e ottimizzazione ha permesso al modello di apprendere in modo efficace pattern complessi, riducendo il rischio di overfitting e migliorando le sue prestazioni su dati non visti.

3.6 Metriche di valutazione

Per una valutazione rigorosa delle prestazioni del modello, sono state utilizzate le seguenti metriche di classificazione, sia durante l'addestramento che nella fase di test finale:

- *True Positives (TP):* Numero di video deepfake correttamente classificati come deepfake.
- *True Negatives (TN):* Numero di video reali correttamente classificati come reali.
- *False Positives (FP):* Numero di video reali erroneamente classificati come deepfake.
- *False Negatives (FN):* Numero di video deepfake erroneamente classificati come reali.

Da queste conteggi di base, vengono derivate le metriche aggregate:

- *Accuracy:* Percentuale di previsioni corrette.
- *Precision:* Precisione delle predizioni positive (deepfake).
- *Recall:* Capacità di individuare tutti i deepfake.
- *F1-Score:* Bilancia precisione e recall, utile con dati sbilanciati.
- *MSE:* Errore medio tra probabilità previste e reali.
- *Log Loss:* Penalizza previsioni poco sicure o errate, misura la qualità delle probabilità.

Queste metriche, interpretate congiuntamente, forniscono una visione completa delle prestazioni del modello, evidenziando sia la sua capacità di identificare i deepfake che di evitare falsi allarmi, cruciale per la sua applicabilità in scenari reali.

4 RISULTATI SPERIMENTALI

In questo capitolo vengono descritti e analizzati i vari approcci adottati per l'attuazione del sistema proposto. L'obiettivo comune di tutti gli esperimenti è stato quello di testare l'architettura CNN-LSTM nel riconoscimento di contenuti deepfake, focalizzandosi sull'analisi della regione labiale.

4.1 Risultati addestramenti

Durante la fase di addestramento del modello, come anticipato nel Capitolo precedente, sono state esplorate diverse configurazioni e strategie, con un approccio incrementale basato sui risultati ottenuti.

4.1.1 Prima sperimentazione: Analisi del Volto Completo. Il primo esperimento ha adottato una configurazione ad alta intensità computazionale, mantenendo in input le immagini dei volti completi con una sequenza temporale di 60 frame per video. Questo approccio, seppur ambizioso, ha causato un rapido esaurimento delle risorse di calcolo disponibili: durante la quarta epoca di addestramento, la GPU e la CPU sono andate incontro a saturazione, rendendo impossibile la prosecuzione del training. Il carico di memoria richiesto per gestire batch di immagini full-frame su sequenze lunghe si è rivelato eccessivo per l'ambiente di sviluppo. Per mitigare i problemi, è stato introdotto un meccanismo di caricamento progressivo dei dati basato sull'uso del modulo tf.data.Dataset, che ha permesso di processare dinamicamente i dati in piccoli batch, riducendo la latenza e la complessità. Un'ulteriore modifica significativa ha riguardato la gestione della memoria: invece di caricare tutte le matrici dei frame in memoria contemporaneamente — soluzione inefficiente sia dal punto di vista della RAM che dei tempi di accesso ai file — si è optato per caricare inizialmente solo i percorsi dei file, suddivisi in input e target, tutto questo, sempre con l'obiettivo di non sovraccaricare la memoria (RAM). Nonostante queste ottimizzazioni, il modello non ha potuto completare l'intero ciclo di apprendimento. Le prime epoche completate hanno mostrato un comportamento gravemente sbilanciato: il modello ha classificato tutti i video reali come deepfake (TN=0). L'accuratezza del 77.1% e l'F1-score di 0.871, sebbene numericamente accettabili, risultavano fuorvianti e non rappresentavano una reale capacità discriminativa, a causa della completa incapacità del modello di distinguere i contenuti reali da quelli manipolati. Questo ha evidenziato la necessità di adottare strategie di bilanciamento o tecniche di regolarizzazione più efficaci.

4.1.2 Seconda sperimentazione: Focalizzazione sulla regione labiale. Nel secondo tentativo si è adottato un approccio più modulare e ottimizzato, intervenendo su due fronti:

- *Riduzione della risoluzione*: Le immagini in input sono state ridotte a 32x32 pixel.
- *Focalizzazione sulla regione labiale*: L'analisi è stata limitata alla sola regione delle labbra. Questa scelta ha consentito di concentrare l'attenzione del modello sui micro-movimenti più rilevanti per il rilevamento delle manipolazioni.
- *Riduzione della sequenza*: La coerenza temporale delle sequenze è stata costituita da 30 frame.

La rete è stata configurata con un numero ridotto di filtri convoluzionali e unità LSTM, bilanciando precisione ed efficienza.

Table 2: Risultati su dataset di addestramento (inclusi esperimenti senza Optical Flow)

Dataset	OF[1]	Accuracy	Precision	Recall	F1-Score	MSE	Log Loss	TP	TN	FP	FN
Celeb	No	0.771	0.771	1.000	0.871	0.229	0.827	1006	0	298	0
Celeb	No	0.784	0.784	1.000	0.879	0.216	0.712	932	0	257	0
Celeb	Sì	0.913	0.921	0.978	0.949	0.087	0.240	814	106	70	18
M2F/WAV2LIP	Sì	0.876	0.890	0.970	0.928	0.124	0.326	2392	226	296	75
Celeb+M2F/WAV2LIP	Sì	0.970	0.970	0.994	0.982	0.030	0.090	1545	302	47	10

[1]OF = Optical Flow

Grazie all'uso della mixed precision e a un batch size ridotto a 4, l'addestramento è stato completato con successo. Tuttavia, il modello mostrava ancora lo stesso problema critico del primo esperimento: una forte tendenza a classificare tutti i campioni reali come deepfake, come evidenziato dall'assenza di veri negativi (TN=0). Il log loss di 0.712, seppur migliorato rispetto al primo tentativo, ha confermato l'incertezza del modello nelle predizioni, evidenziando la necessità di migliorare drasticamente la capacità discriminativa tra le classi.

4.1.3 Terza sperimentazione: Selezione dei frame tramite Optical Flow. In questo test si è deciso di introdurre una tecnica di selezione dinamica dei frame basata sull'Optical Flow. L'obiettivo è stato selezionare i frame che presentano la maggiore attività di movimento, in quanto sono più propensi a contenere informazioni critiche per rilevare anomalie. I frame sono stati selezionati utilizzando una soglia di intensità del movimento pari a 1.0, garantendo una sequenza più significativa. Inoltre, è stato possibile tornare a utilizzare 60 frame (selezionati in base al movimento, focalizzandosi sempre sulla regione delle labbra come ROI principale), beneficiando così di un contesto temporale più ampio senza compromettere le risorse di calcolo. L'effetto di queste scelte si è rivelato decisivo: il modello ha ottenuto risultati realmente bilanciati con un'accuratezza del 91.3% e un F1-score di 0.949, con un recall del 97.8% e, soprattutto, valori di True Negative diversi da zero (TN=106). Questo ha dimostrato la capacità di riconoscere correttamente i video autentici, superando il problema di classificazione sbilanciata. La riduzione del log loss a 0.240 e del MSE a 0.087 ha confermato l'effetto positivo dell'inclusione dell'Optical Flow nella pipeline di pre-processing. A partire da questo punto, tutti gli addestramenti successivi fatti sul dataset M2Fred/WAV2LIP e sulla sua fusione con Celeb-DFv2 sono stati condotti adottando tale tecnica di selezione dei frame basata sull'Optical Flow, che si è dimostrata determinante per ottenere prestazioni più robuste e generalizzabili. La Tabella 2 riassume i risultati di addestramento, evidenziando il miglioramento progressivo con l'introduzione dell'Optical Flow.

4.2 Risultati Test Cross-Dataset

Per valutare la capacità di generalizzazione dei modelli addestrati, sono stati condotti test su dataset non utilizzati durante l'addestramento. Questo tipo di test è fondamentale per verificare l'effettiva robustezza dei modelli nel riconoscimento di manipolazioni prodotte con tecniche diverse o su domini visivi differenti. La Tabella 3 riassume i risultati di questi test.

- *Modello addestrato su Celeb-DFv2, testato su M2Fred/WAV2LIP*: Il modello ha evidenziato un'elevata sensibilità (recall perfetto, 1.0) e un numero totale di falsi negativi pari a zero. Tuttavia, la precisione si è fermata a 0.808 a causa della presenza di falsi positivi (573 FP). L'accuratezza complessiva è risultata pari a 80.8%, con un F1-score di 0.894 e un log loss pari a 0.470. Questo comportamento suggerisce che il modello addestrato su Celeb-DFv2 tende a sovrastimare la presenza di manipolazioni quando si trova di fronte a tecniche di generazione differenti da quelle viste in fase di addestramento.
- *Modello addestrato su Celeb-DFv2, testato su XM2VTS*: Quando lo stesso modello è stato testato su XM2VTS, un dataset composto esclusivamente da video reali, ha classificato correttamente tutte le istanze, ottenendo un'accuratezza del 100% e un errore quadratico medio nullo. Tuttavia, l'assenza di video deepfake ha impedito il calcolo di metriche quali precision, recall e F1-score, in quanto non vi erano casi positivi su cui eseguire tali valutazioni.
- *Modello addestrato su M2Fred/WAV2LIP, testato su Celeb-DFv2*: Anche in questo caso si è riscontrata un'ottima capacità di individuare i contenuti falsi (recall = 0.995), ma la precisione si è mantenuta moderata (0.826) a causa della presenza di 580 falsi positivi. Il valore di F1-score è stato pari a 0.903, con un'accuratezza dell'82.3% e un log loss di 0.473. Questo risultato è analogo a quello ottenuto nel test inverso, confermando una certa simmetria tra i due modelli in termini di prestazioni cross-dataset.
- *Modello addestrato su M2Fred/WAV2LIP, testato su XM2VTS*: Il modello ha mostrato un'accuratezza del 99.2%, con un MSE molto basso (0.008). Anche in questo caso, la totale assenza di contenuti manipolati nel dataset ha impedito il calcolo delle metriche di classificazione più comuni, ma l'alto numero di veri negativi identificati (1192 su 1202) dimostra la solidità del modello nel riconoscere video autentici.
- *Modello Combinato (addestrato su Celeb-DFv2 + M2Fred/WAV2LIP), testato su XM2VTS*: Infine, è stato effettuato un test sul dataset XM2VTS utilizzando il modello addestrato sui set combinati.

Table 3: Risultati dei test cross-dataset

Model	Test Set	Accuracy	Precision	Recall	F1-Score	MSE	Log Loss	TP	TN	FP	FN
Celeb	M2/WAV2	0.808	0.808	1.000	0.894	0.192	0.470	2412	4	573	0
Celeb	XM2	1.000	–	–	–	0.000	–	0	1202	0	0
M2/WAV2	Celeb	0.823	0.826	0.995	0.903	0.177	0.473	2758	6	580	13
M2/WAV2	XM2	0.992	–	–	–	0.008	–	0	1192	10	0
Celeb+M2F/WAV2	XM2	1.000	–	–	–	0.000	–	0	1202	0	0

Anche in questo caso, il modello ha classificato correttamente tutte le istanze reali, raggiungendo il 100% di accuratezza e un MSE nullo.

Questi risultati mostrano chiaramente che i modelli addestrati su singoli dataset sono in grado di ottenere ottime prestazioni nel contesto specifico in cui sono stati addestrati. Tuttavia, il modello combinato si distingue per un comportamento più bilanciato e robusto rispetto alla variabilità cross-dataset, confermandosi come la soluzione più adatta per applicazioni nel mondo reale grazie alla sua maggiore adattabilità e versatilità.

5 CONCLUSIONI

Il presente lavoro ha proposto una soluzione al problema del rilevamento automatico dei contenuti deepfake, basata su un'architettura ibrida CNN-LSTM. Il progetto si è focalizzato sulla regione labiale e si è sviluppato attraverso un processo iterativo di progettazione, ottimizzazione e valutazione su differenti dataset, con l'obiettivo di realizzare un sistema robusto e generalizzabile.

Le analisi condotte dimostrano che, se opportunamente ottimizzato, il modello CNN-LSTM è efficace nell'apprendere le caratteristiche salienti che distinguono contenuti autentici da quelli manipolati. In particolare, l'introduzione di tecniche chiave quali l'Optical Flow per la selezione dei frame più significativi, la focalizzazione sulla zona delle labbra, l'uso della mixed precision e un adeguato schema di regolarizzazione si sono rivelate decisive per il miglioramento delle performance complessive del sistema.

L'addestramento su dataset eterogenei ha portato a risultati particolarmente significativi, permettendo di superare i limiti di generalizzazione dei modelli addestrati su singoli domini. Il modello combinato ha dimostrato prestazioni elevate sia in scenari interni che su dati completamente esterni come quelli del dataset XM2VTS, classificando tutti i video reali in modo corretto.

Inoltre, sebbene i dataset presentassero un evidente sbilanciamento tra le classi, questo non si è rivelato un problema significativo nel processo di addestramento. Le prove con l'uso di pesi per bilanciare le classi non hanno mostrato una necessità sistematica di tali strategie, grazie anche alla distribuzione proporzionata tra i set di training e validation, che ha contribuito a mantenere l'equilibrio nelle prestazioni del modello.

Sviluppi futuri

I risultati conseguiti aprono la strada a molteplici direzioni di ricerca e sviluppo per l'evoluzione del sistema:

- *Integrazione Multimodale:* sviluppare un approccio che combini informazioni audio e video per identificare incongruenze tra movimento labiale e contenuto vocale, potenziando significativamente la capacità di rilevare manipolazioni di nuova generazione.
- *Espansione e Diversificazione dei Dataset:* arricchire i training set con dataset eterogenei di deepfake, includendo tecniche di manipolazione emergenti e contenuti provenienti da diverse fonti, al fine di incrementare la robustezza e la capacità di generalizzazione del modello.
- *Implementazione Real-Time:* ottimizzare l'architettura computazionale per consentire l'analisi di flussi video in tempo reale, abilitando applicazioni critiche come la moderazione

automatica di contenuti e la verifica istantanea di trasmissioni live.

- *Specializzazione Domain-Specific:* implementare strategie di transfer learning (riuso di modelli pre-addestrati) e fine-tuning adattivo (ottimizzazione mirata su nuovi dati) per ottimizzare le prestazioni in contesti applicativi specifici, quali piattaforme social media, ambienti corporate o verifiche giornalistiche.
- *Interpretabilità e Trasparenza:* integrare tecniche di Explainable AI (XAI) per garantire trasparenza decisionale, fornendo visualizzazioni che evidenzino le regioni dell'immagine maggiormente influenti nella classificazione e migliorando la fiducia degli utenti nel sistema.

REFERENCES

[1] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Younsik Shin, and Simon S. Woo. 2020. Clrnet: a deep convolutional lstm residual network for deepfake video detection.

[2] Trisha Mitra, Piyush Banerjee, Shubham Roy, Indranil Rampalli, and Arnab Chatterjee. 2020. Deepfake detection using spatiotemporal convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] Johnson B. Awotunde, Rasheed G. Jimoh, Olatunji M. Odeniyi, Abdulrahaman O. Ameen, Mary K. Abiodun, and Emmanuel A. Adeniyi. 2022. A deep learning-based deepfake detection and classification model using convolutional neural network. *IEEE Access*.

[4] Irene Amerini, Leonardo Galteri, Alberto Del Bimbo, Nicholas Clemencic, and Saptarshi Sarkar. 2023. Poiforensics: person-of-interest based audio-visual deepfake detection. In.

[5] Liming Wang, Haisheng Wang, Xu Zhang, Bo Su, and Ming Yang. 2020. Deepfake video detection based on motion prediction patterns. In.

[6] Junhwa Hur and Stefan Roth. 2020. Optical flow estimation in the deep learning age. *arXiv preprint arXiv:2004.02853*. https://arxiv.org/abs/2004.02853.