Masked Face Recognition tramite Facial Dynamics

Alessandro Ferrentino
Università degli Studi di Salerno
Dipartimento di Informatica

Abstract

L'esplosione della pandemia COVID-19 ha favorito lo sviluppo e la diffusione di sistemi biometrici basati sul riconoscimento facciale ma ha reso l'utilizzo della mascherina una costante in tutto il mondo. Lo scopo del seguente lavoro è quello di risolvere il problema del Masked Face Recognition mediante Facial Dynamics, utilizzando quindi come riferimento sia sequenze video di soggetti con indosso una mascherina che senza. La strategia risolutiva si fonda sull'idea dell'utilizzare le regioni del volto poco affette dalla presenza della mascherina e quindi sull'estrazione della regione perioculare. Successivamente, si applica un'architettura CNN-LSTM sulle sequenze contenenti la sola zona di interesse al fine di utilizzare le dinamiche facciali per il processo di classificazione. In fine, si conducono delle sperimentazioni per analizzare il comportamento della rete sui dati a disposizione, in virtù della soluzione impiegata.

1 INTRODUZIONE

Un sistema di riconoscimento biometrico è un particolare tipo di sistema informatico che ha la funzionalità e lo scopo di identificare una persona sulla base di una o più caratteristiche fisiologiche (e.g. impronta digitale, iride, volto) e/o comportamentali (e.g. voce, stile di battitura, firma), chiamate caratteristiche biometriche.

Le metodologie di acquisizione di queste caratteristiche sono intrinsecamente legate alla natura delle stesse e si dividono in due categorie: per contatto e senza contatto (contactless).

Il volto è una caratteristica biometrica contactless, ampiamente utilizzata nei sistemi di controllo degli accessi e nei sistemi di videosorveglianza, soprattutto grazie alla metodologia con cui può essere acquisita ed al fatto che non necessita di cooperazione da parte del soggetto. Inoltre, grazie ai progressi nell'ambito delle tecnologie informatiche, la diffusione di sistemi di riconoscimento facciale si è estesa in modo capillare anche ai dispositivi mobile ed embedded [1], trovando così applicazione nei sistemi di autenticazione mobile (e.g. Face ID).

Con l'esplosione della pandemia COVID-19, gli esperti hanno caldamente suggerito di evitare qualsivoglia tipologia di contatto fisico e di indossare la mascherina, al fine di contrastare la diffusione del virus. Questo ha posto dunque un serio vincolo all'applicazione di sistemi basati su caratteristiche biometriche da contatto (e.g. impronta digitale) in scenari pubblici, quali ad esempio il controllo di frontiera, a favore di sistemi basati su caratteristiche contactless, come il volto.

In virtù di ciò, osserviamo che l'avvento del COVID ha ulteriormente favorito l'utilizzo di sistemi basati sul riconoscimento facciale ma, d'altra parte, ha introdotto con enorme frequenza l'utilizzo della mascherina, elemento di grande disturbo per quanto concerne l'estrazione delle caratteristiche associate al volto e che minaccia dunque l'efficacia di questi sistemi.

Lo scopo di questo progetto è quello di risolvere il task del riconoscimento facciale con mascherina tramite le dinamiche facciali, ovvero sequenze temporali delle caratteristiche del volto. A tal fine saranno usate, come riferimento, sia sequenze video di soggetti con indosso una mascherina, che senza, e verranno adoperate una serie di tecniche e tecnologie di image processing, computer vision e machine learning.

2 RELATED WORKS

Le occlusioni sono un'importante limitazione nel contesto del riconoscimento facciale. In genere queste sono causate dal fatto che i soggetti indossano cappelli, occhiali, mascherine o, ad ogni modo, qualsiasi oggetto che copre parzialmente il volto. Dunque, indossare la mascherina è considerata la sfida più grande per quanto concerne le occlusioni parziali del volto, dal momento che nasconde gran parte del viso, incluso il naso. Nella ricerca, sono state sviluppate diverse tecniche che permettono di trattare questo genere di problema. In particolare faremo riferimento alle tecniche di rimozione delle occlusioni.

Questo genere di tecniche puntano a rilevare le zone del volto occluse e a rimuoverle completamente dal processo di estrazione delle caratteristiche e di classificazione. Uno degli approcci migliori è quello basato sulla segmentazione, in cui dapprima viene rilevata la regione occlusa e successivamente viene utilizzata solo la parte non occlusa per gli step successivi.

Per esempio, Priya e Banu [2] hanno diviso il volto in piccole zone locali. Successivamente, per rimuovere la regione occlusa, hanno applicato un classificatore SVM (Support Vector Machine) per rilevarla. In fine, per eseguire il face recognition, viene utilizzata una matrice con pesi basati sulla media sulle regioni non occluse.

Dalla pubblicazione dell'architettura AlexNet nel 2012 da parte di Krizhevsky et al. [3], le deep CNN sono diventate un approccio comune nel face recognition. Queste reti sono state usate con successo anche nella variante con occlusione del face recognition [4]. E' stato osservato che il metodo di apprendimento della rete si basa sul fatto che il sistema di visione umano ignora automaticamente le regioni occluse del volto e si concentra esclusivamente su quelle non coperte. Per esempio, Song et al. [5] hanno proposto una tecnica di apprendimento della mascherina in modo tale da rimuovere le caratteristiche associate alle regioni della mascherina nel processo di riconoscimento.

Tutti questi lavori tuttavia usano come input singole immagini,

e quindi non permettono di modellare la struttura temporale in sequenze di immagini. Per estrarre le caratteristiche temporali da sequenze video, Bac-couche et al. [6] applicano le unità LSTM in cascata alle caratteristiche estratte con un algoritmo Scale-Invariant Feature Transform (SIFT). Applicando quindi questa strategia ad una struttura CNN, si ottiene un modello CNN-LSTM, in cui le unità LSTM sono poste in cascata alla CNN.

3 SISTEMA PROPOSTO

Prima di passare alla presentazione della strategia utilizzata nel risolvere il problema del masked face recognition, si fornisce una descrizione dei dataset di riferimento contenenti le sequenze video necessarie per lo sviluppo della soluzione.

M2FRED

M2FRED (Mobile Masked Face REcognition Database) è un dataset sviluppato dal BIPLab, presso l'Università degli Studi di Salerno, contenente sequenze video di soggetti di etnia prevalentemente mediterranea, sesso sia maschile che femminile, e la cui età spazia tra quella del giovane adulto (circa 25 anni) ed età adulta media (40-60 anni), con una preponderanza della prima categoria.

Le acquisizioni delle sequenze, per ogni soggetto, sono avvenute in 4 differenti sessioni, intervallate da pochi giorni, e alternando sessioni indoor e sessioni outdoor, per un totale di 2 sessioni indoor e 2 sessioni outdoor. Ogni sessione è caratterizzata da 4 acquisizioni con mascherina e 4 senza mascherina. Durante l'acquisizione di ciascuna sequenza, il soggetto si pone frontalmente alla camera e pronuncia una breve frase dalla durata di pochi secondi. Nel corso delle acquisizioni, al soggetto è permesso un certo grado di libertà, consentendogli di porsi ad una distanza variabile dalla camera, e ad assumere pose non completamente frontali.

In virtù della libertà concessa ai soggetti ed al fatto che le acquisizioni avvengono in ambienti non controllati, le sequenze video sono caratterizzate da una forte variabilità in termini di luminosità, posa e distanza rispetto alla camera.

Complessivamente il dataset consta di 46 soggetti. Ad ogni soggetto sono associati 16 video con mascherina e 16 video senza mascherina. Le caratteristiche dei video (risoluzione ed fps) variano tra i soggetti, dal momento che sono state impiegate diverse camere d'acquisizione, tuttavia tutte le sequenze sono contraddistinte da una breve durata (< 12 secondi).

Figure 1: M2FRED

XM2VTS

XM2VTS è un dataset sviluppato dell'Università di Surrey contenente sequenze video di soggetti di differente etnia, sesso sia maschile che femminile, e la cui età spazia tra quella del giovane adulto ed età adulta media.

Le acquisizioni delle sequenze, per ogni soggetto, sono avvenute nello stesso intervallo temporale, per un totale di 15 video: 11 in cui il soggetto è posto in modo frontale rispetto alla camera e pronuncia una breve frase di pochi secondi, e 4 in cui il soggetto esegue diverse rotazioni con la testa, portando il volto di profilo, verso l'alto e verso il basso. Durante le acquisizioni, il soggetto è posto ad una precisa distanza dalla camera, non esegue variazioni di posa indesiderate e l'ambiente è fortemente controllato, garantendo così una buona illuminazione.

Complessivamente il dataset consta di 294 soggetti. Ad ogni soggetto sono associati 15 video, tutti senza mascherina. I video presentano una risoluzione di 720x576 pixel, 25 frames/secondo ed hanno tutti una durata breve (< 12 secondi).

Figure 2: XM2VTS

Osservazioni relative ad XM2VTS

A seguito dell'analisi effettuata sui dataset, possiamo compiere due importanti osservazioni relative ad XM2VTS:

(1) Il dataset contiene video con rotazioni.
(2) All'interno del dataset non sono presenti sequenze con mascherina.

Per quanto concerne il primo punto, dal momento che si ha interesse nell'analizzare le dinamiche facciali relative a soggetti in posizione frontale, le suddette sequenze non risultano pertinenti per il raggiungimento dello scopo preposto. Per tale ragione si è deciso di non considerare queste sequenze.

Figure 3: Rotazione verso sinistra

Relativamente al secondo punto, poiché è di vitale importanza avere a disposizione materiale in cui i soggetti indossano la mascherina, è stato deciso di ricorrere a degli strumenti che permettono di sovrapporne una (fittizia) al volto dei summenzionati. Considerato che in M2FRED il numero di sequenze con mascherina pareggia

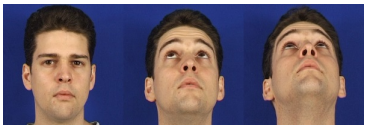

Figure 4: Rotazione verso l'alto

quello senza mascherina, è stato deciso di mantenere lo stesso rapporto anche in XM2VTS, mantenendo così 5 sequenze intatte e sovrapponendo la mascherina ad altre 5 sequenze.

Applicazione della mascherina

Per applicare le mascherine ai soggetti è stato fatto uso di un tool open-source, scritto in linguaggio python, chiamato face-mask [7]. Il tool in esame utilizza innanzitutto un algoritmo di face detection sull'immagine ricevuta in input, in modo tale da individuare il volto del soggetto. Successivamente, provvede ad apporre la mascherina sul volto del soggetto, cercando di adattarne le dimensioni in modo tale da coprire la zona del mento e, in modo parziale, gli zigomi ed il naso, mimando quindi una mascherina reale.

Dal momento che si hanno a disposizione sequenze video ma il tool opera su immagini, è stato fatto uso della libreria OpenCV per estrarre i frame da ciascun video, così da poter impiegare il tool su ognuno di essi, ed ottenere quindi le sequenze di frame con mascherina.

Figure 5: Mascherina apposta ad un soggetto

Soluzione proposta

Per poter eseguire il riconoscimento facciale dei soggetti presenti nei dataset appena descritti, dal momento che in alcune sequenze questi indossano la mascherina, è stata considerata una strategia complementare alle tecniche di rimozione delle occlusioni: anziché rimuovere la zona occlusa così da lavorare solo su zone non occluse, si vuole estrarre direttamente la zona che risente di meno degli effetti della mascherina, ovvero la zona perioculare.

Inoltre, si vuole utilizzare una soluzione basata sul machine learning che permetta di trattare sequenze di immagini. Per questa ragione è stato proposto l'impiego di una particolare deep neural network chiamata CNN2D-LSTM, che permette di estrarre sia le caratteristiche spaziali da ciascun frame, che le caratteristiche temporali tra i frames.

Possiamo quindi individuare due fasi fondamentali nella realizzazione della soluzione:

(1) Una fase di preprocessing, nella quale i video in input vengono elaborati in modo tale da ottenere sequenze di frames della regione perioculare, ed in cui i dati vengono uniformati.

(2) Una fase di estrazione delle caratteristiche e di classificazione, così da avere una struttura che ci permetta di riconoscere il volto dei soggetti.

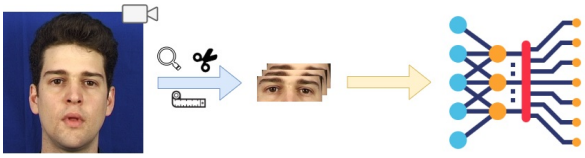

Figure 6: Workflow della strategia applicata

3.1 Fase di Preprocessing

Per poter eseguire l'estrazione della zona perioculare è necessario innanzitutto riuscire ad individuarla all'interno dell'immagine. A tale scopo è stato fatto uso della libreria dlib, una libreria di machine learning avanzata, adottata principalmente per rilevare i punti di riferimento (landmarks) facciali di un individuo. Il rilevamento dei landmark facciali è definita come un'attività nel rilevare i punti di riferimento chiave sul viso (occhi, sopracciglia, naso, mascella, bocca...) e nel tracciarli, ossia determinare le loro coordinate spaziali. Per poter individuare i suddetti è necessario innanzitutto rilevare il viso, mediante un face detector, e successivamente adoperare un predictor sul volto trovato, così da ottenere le posizioni.

In particolare, per quanto riguarda il detector, sono stati utilizzati, in chiave comparativa, sia quello basato sulla tecnica HOG+SVM che quello basato su CNN, entrambi contenuti in dlib. Per quanto concerne il predictor, è stato impiegato un predictor che localizza 68 landmark facciali, anche questo incluso in dlib.

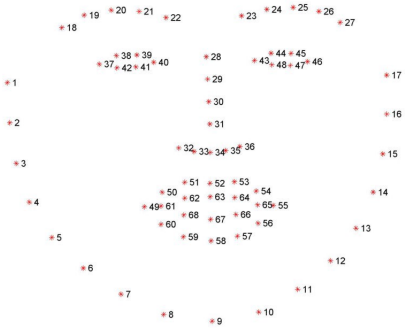

Figure 7: 68 landmark facciali

Una volta localizzati i punti di riferimento, per poter estrarre la zona di interesse, sono stati utilizzati, in riferimento alla figura 7, come riferimenti verticali (sinistro e destro), i landmark in posizione 1 e 17 rispettivamente. Relativamente ai riferimenti orizzontali, viene individuato il landmark localizzato più in alto (tra quelli associati alle sopracciglia) ed il landmark localizzato più in basso (tra quelli associati alla parte inferiore degli occhi e quello in posizione 29). Il motivo per cui i riferimenti orizzontali vengono individuati in modo dinamico piuttosto che statico, è legato al fatto che i soggetti non assumono sempre pose dritte ma talvolta inclinano la testa verso una direzione, il che porterebbe ad un taglio errato della

regione di interesse. Una volta localizzati i 4 riferimenti, questi descrivono quindi un rettangolo che delimita la regione da tagliare: la zona perioculare.

3.1.1 *Preprocessing M2FRED.*

Relativamente al dataset M2FRED, è stata applicata la strategia appena descritta tramite due approcci: in uno è stato usato il detector HOG, nell'altro il detector CNN. Comparando il numero di frames ricavati, osserviamo che l'approccio basato sulla CNN produce circa il 40% in più di frames rispetto all'approccio HOG, risultando quindi significativamente più efficace.

In particolare, analizzando i risultati ottenuti, si osserva che circa l'89% delle sequenze, presenta un numero di frames superiore a 50; mentre le restanti contano un numero di frames inferiore. Inoltre, poiché nel dataset sono presenti sequenze video con un numero massimo di 50 frames, è stato posto l'obiettivo di riuscire ad ottenere almeno 50 frames da ciascuna sequenza, e si vuole quindi trovare una strategia per raggiungere questo obiettivo sul restante 11% dei dati.

Dall'analisi del dataset, emerge che alcuni video sono caratterizzati da soggetti che presentano forti occlusioni, dal momento che indossano sia la mascherina che gli occhiali; altri invece sono caratterizzati da una cattiva illuminazione. In virtù di ciò, è stata proposta dunque una strategia basata sul contrast enhancement, al fine di migliorare il contrasto delle sequenze relativamente alle quali non è stato possibile ottenere il numero di frames convenuto, ed agevolare così il processo di detection.

In questo contesto, sono state utilizzate, in chiave comparativa, due tecniche:

(1) La tecnica dell'equalizzazione dell'istogramma. Lo scopo fondamentale di questa tecnica è quello di migliorare il contrasto dell'immagine, operando sull'istogramma correlato ad essa, così che ad ogni livello di grigio sia associato circa lo stesso numero di pixel.
(2) Una particolare tecnica di equalizzazione denominata CLAHE (Contrast Limited Adaptive Histogram Equalization). Così come l'equalizzazione dell'istogramma, anche questa tecnica punta a migliorare il contrasto delle immagini ma anziché operare sull'istogramma dell'intera immagine, esegue un'operazione di equalizzazione sui vari istogrammi associati a delle porzioni dell'immagine.

I frame così processati vengono successivamente trattati con un detector CNN, in riferimento alla strategia precedentemente discussa, per l'estrazione della zona di interesse. Per quanto riguarda la prima, è stato possibile recuperare circa il 20% delle sequenze non conformi (con meno di 50 frames). Relativamente alla seconda, ne sono state salvate il 38.4%.

Dal momento che risulta cruciale riuscire ad ottenere quante più sequenze possibili, date le dimensioni del dataset, sulle restanti sequenze è stato adoperato un taglio statico della zona di interesse. In particolare, è stato utilizzato un breve script in python che, date in input le coordinate della zona pericoulare, provvede ad effettuare il taglio della zona desiderata.

Ciononostante, dal momento che 3 sequenze di 3 differenti soggetti contano un numero massimo di 25 frames, non è stato possibile ottenere 50 frames da tutte quelle presenti nel dataset. Questa considerazione porta a 3 possibili soluzioni:

(1) Considerare solo 25 frames per ogni sequenza.
(2) Escludere i soggetti dal dataset processato.
(3) Escludere le sequenze problematiche dal dataset processato.

Tra queste è stato deciso di optare per la terza strategia, dal momento che è quella che permette di avere a disposizione il maggior numero di soggetti, mantenendo comunque un buon numero di sequenze per ciascuno di essi.

Giacché risulta fondamentale uniformare i dati per poterli trattare nei passi successivi, considerando che alcuni soggetti presentano solo 15 sequenze con o senza mascherina, è stata eliminata per ogni soggetto, ove necessario, 1 sequenza con mascherina ed 1 senza mascherina. Inoltre ciascuna sequenza è stata troncata ad un numero esatto di 50 frames, e ciascun frame è stato riadattato ad una dimensione di 200x200 pixel.

Nella tabella 1 si sintetizzano i dati relativi ad M2FRED al termine dell'operazione di preprocessing.

Num. Soggetti	Seq. con mascherina	Seq. senza mascherina	Num. Frames	Dim. Frame
46	15	15	50	200x200 px

Table 1: M2FRED

3.1.2 *Preprocessing XM2VTS.*

Visto che in M2FRED è stato stabilito un numero di 50 frames per ciascuna sequenza, si vogliono ottenere sequenze della stessa lunghezza in XM2VTS, così da avere come unici elementi di variabilità tra i dataset (in termini di dimensionalità dei dati), il numero di soggetti ed il numero di sequenze associate a ciascuno di essi.

Sulla base dei risultati ottenuti tramite gli approcci impiegati in M2FRED, in XM2VTS è stata applicata la stessa strategia, utilizzando il detector CNN. Questo ha permesso di ottenere, relativamente alle sequenze senza mascherina, 50 frames per ognuna di esse; mentre per quanto riguarda quelle con mascherina, non è stato possibile ottenere 50 frames per alcune di queste, a causa della presenza dell'occlusione. Per risolvere il problema, ed evitare di adottare nuovamente una soluzione basata sul taglio statico, è stata proposta la seguente strategia:

(1) Viene eseguita la face detection sulle sequenze video originali (prima che venisse applicata la mascherina), così da ottenere le posizioni relative ai landmark che individuano la zona perioculare
(2) Si utilizzano i landmark per eseguire il taglio della regione di interesse nelle sequenze con mascherina

La ratio alla base della seguente strategia risiede nel fatto che la posizione della regione perioculare è invariata nella sequenza con mascherina rispetto all'originale. Ad ogni modo, procedendo nel modo descritto, è stato possibile in fine ottenere 50 frames su tutte

le sequenze con mascherina.

Per quanto concerne l'uniformazione dei dati, anche in questo caso è stato stabilito di ridimensionare i frame a 200x200 pixel. Nella tabella 2 si sintetizzano i dati relativi ad XM2VTS al termine dell'operazione di preprocessing.

Num. Soggetti	Seq. con mascherina	Seq. senza mascherina	Num. Frames	Dim. Frame
294	5	5	50	200x200 px

Table 2: XM2VTS

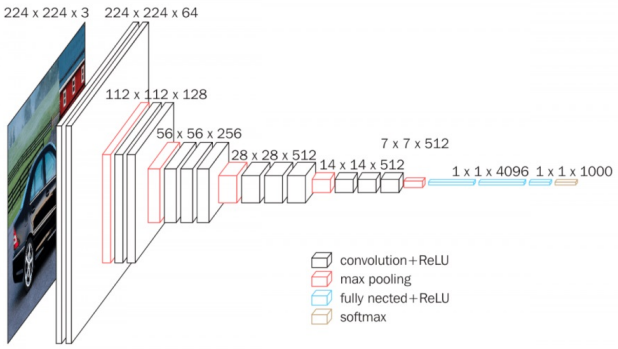

Figure 9: Architettura della rete VGG16

3.2 Rete Neurale

Per effettuare il processo di classificazione, e riconoscere quindi i soggetti, è stata utilizzata un'architettura CNN2D-LSTM. Questa è costituita da 3 elementi fondamentali:

(1) Una struttura CNN2D, la quale, ricevuta in input un'immagine, ne estrae le caratteristiche spaziali e le cede al layer successivo.
(2) Una struttura LSTM, incaricata di estrarre le caratteristiche temporali associate alla sequenza di caratteristiche precedentemente estratte dalla CNN. Le caratteristiche così ottenute vengono quindi inviate al layer successivo.
(3) Uno strato fully connected, che ha lo scopo di generalizzare dalle caratteristiche ricevute in input, dando come risposta un elemento appartenente allo spazio di output, e quindi un'etichetta che rappresenta l'identità del soggetto.

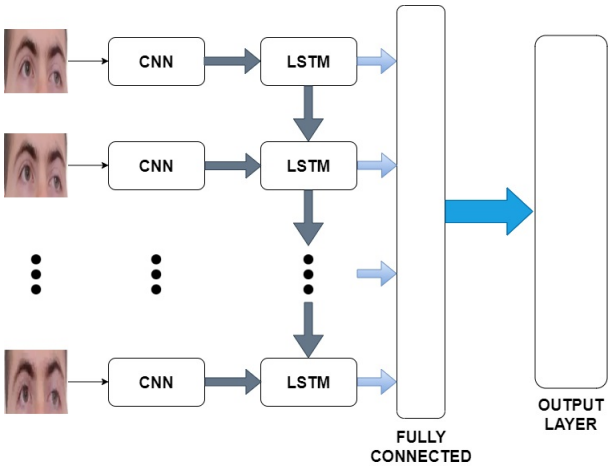

Figure 8: CNN-LSTM spiegata nel tempo

In particolare, l'architettura di riferimento [8] utilizza una peculiare CNN nota come VGG16 (figura 9), ampiamente usata nell'estrazione delle caratteristiche relative alle immagini, dalla quale sono stati rimossi gli ultimi layer (fully nected+ReLU e softmax). In cascata alla CNN è posto inoltre un layer di flatten, in modo tale da ridurre la dimensionalità delle caratteristiche. Seguono quindi lo strato LSTM e Fully Connected.

Nel dettaglio, sono stati considerati due modelli basati sull'architettura appena presentata:

(1) Un modello in cui la rete VGG16 è stata pre-addestrata sul dataset Imagenet, un enorme database formato da più di 14 milioni di immagini di vario genere.
(2) Un modello in cui la rete VGG16 è stata pre-addestrata sul dataset VGGFace, un database formato da 3.31 milioni di immagini di 9131 soggetti ottenute tramite Google Image Search.

4 RISULTATI SPERIMENTALI

In relazione all'architettura presentata, sono stati eseguiti vari esperimenti per testare quale fosse il migliore modello, la migliore configurazione dei parametri e per studiare i dati contenuti nei dataset inizialmente introdotti.

4.1 Modello 1 vs Modello 2

La prima sperimentazione condotta ha l'obiettivo di valutare quale tra i due modelli considerati offre le migliori prestazioni in termini di accuratezza. Nell'esecuzione delle sperimentazioni è stato effettuato il freezing dei layers della struttura VGG16. Per quanto concerne i parametri, è stata utilizzata la seguente configurazione:

- **No. unità LSTM:** 512
- **batch size:** 4
- **learning rate:** 0.0001

Come dataset è stato fatto riferimento ad M2FRED (Tabella 1). In particolare è stato fatto uso delle sequenze senza mascherina come training set (50% dei dati totali), e delle sequenza con mascherina come test set (restante 50%).

In Figura 10 sono riportati i risultati ottenuti con il modello preaddestrato sul dataset Imagenet, mentre in Figura 11 sono riportati quelli relativi al modello preaddestrato sul dataset VGGFace. In tabella 3 è possibile osservare i più alti livelli di accuratezza raggiunti dai due modelli.

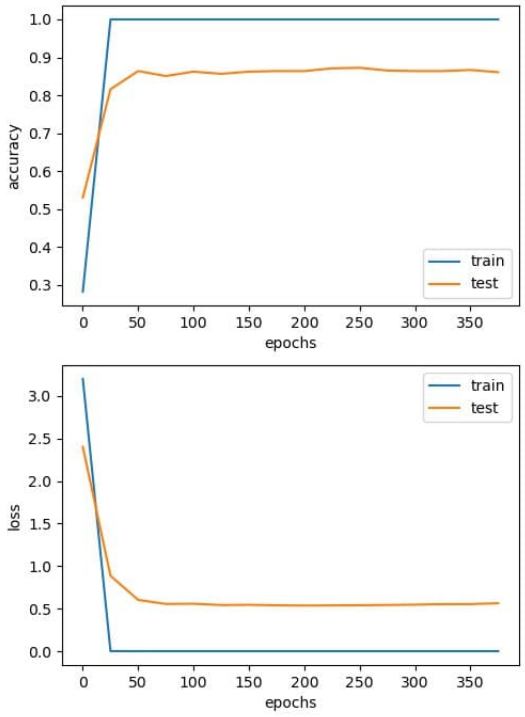

Figure 10: Risultati ottenuti sul modello preaddestrato con Imagenet.

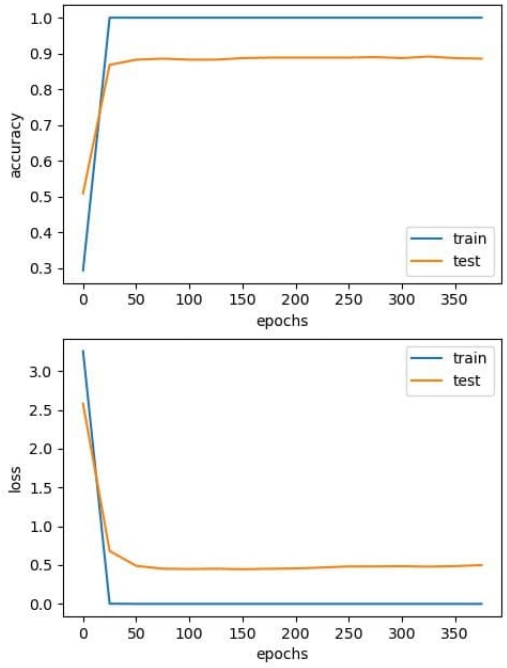

Figure 11: Risultati ottenuti sul modello preaddestrato con VGGFace.

	Imagenet	VGGFace
Accuracy	87.25%	89.13%

Table 3: Massimi valori di accuracy ottenuti con i due modelli

4.2 Tuning degli iperparametri

E' stato eseguito il tuning degli iperparametri in relazione al Modello 2 (VGG16 pre-trained con VGGFace) con freezing dei layers relativi alla struttura CNN. In particolare sono state provate diverse combinazioni:

- **No. unità LSTM:** 64, 128, 256, 512
- **batch size:** 4
- **learning rate:** 0.001, 0.0001, 0.00001
- **epoche:** variabili

Sfortunatamente non è stato possibile provare diversi valori di batch size a causa delle limitazioni in termini di risorse hardware. Il numero di epoche è stato aumentato progressivamente fino ad individuare un andamento stabile o decrescente nei valori di accuracy. Come dataset di riferimento è stato fatto uso di M2FRED (Tabella 1), utilizzando le sequenze senza mascherina come training set e le sequenza con mascherina come test set.

Nella Tabella 4, sono riportati i massimi valori di accuracy raggiunti con le varie combinazioni di parametri.

	No. unità LSTM			
learning rate	64	128	256	512
0.001	19.85%	45.99%	66.66%	67.10%
0.0001	86.37%	88.40%	88.84%	89.13%
0.00001	69.56%	80.29%	82.90%	81.45%

Table 4: Massimi valori di accuracy al variare della learning rate e numero di unità LSTM

4.3 Unfreezing dei layers

E' stata eseguita un'analisi comparativa dei risultati per osservare le conseguenze prestazionali dell'unfreezing dei layers della struttura VGG16. In particolare è stato confrontato il valore di accuracy ottenuto dal Modello 2 con freezed layers e quello conseguito dal Modello 2 con layers unfreezed. Nell'esecuzione delle sperimentazioni sono stati utilizzati i seguenti parametri:

- **No. unità LSTM:** 512
- **batch size:** 4
- **learning rate:** 0.0001

Come dataset ci si è avvalsi di M2FRED, impiegando le sequenze senza mascherina come training set e le sequenze con mascherina come test set.

In Figura 12 è possibile osservare l'andamento dei valori di loss ed accuracy al variare delle epoche relativo al Modello con

layers unfreezed. In Figura 11 si possono esaminare i medesimi relativi al Modello con layers freezed. In Tabella 5 sono riportati i massimi valori di accuracy ottenuti.

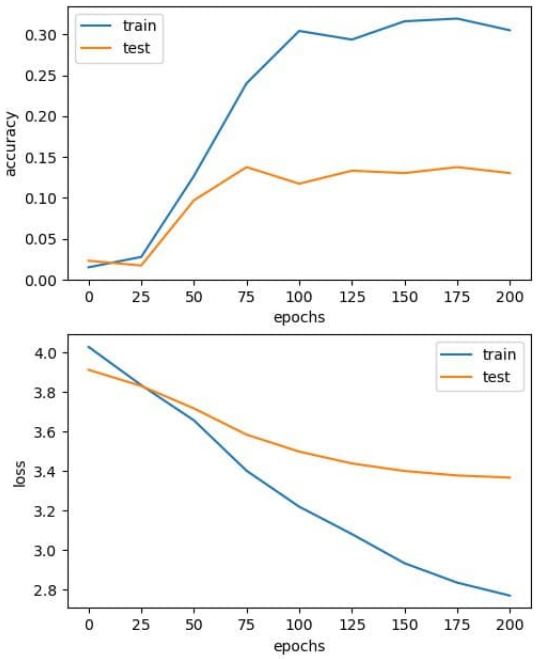

Figure 12: Risultati ottenuti sul modello preaddestrato con VGGFace ed unfreezing dei layers.

	Freezed Layers	Unfreezed Layers
Accuracy	89.13%	13.76%

Table 5: Massimi valori di accuracy ottenuti con freezing/unfreezing dei layers

4.4 Sperimentazioni relative ai dataset

Sono state condotte delle sperimentazioni per analizzare l'impatto della mascherina nel processo di training e le differenze prestazionali sia al variare della quantità dei dati che nell'impiego di sequenze di qualità inferiore (cattiva illuminazione e posa) rispetto a sequenze di qualità superiore.

Nello specifico sono state definite 3 tipologie di esperimenti:

(1) **Tipo 1:** Vengono selezionate le sequenze senza mascherina come training set e le sequenze con mascherina come test set.
(2) **Tipo 2:** Vengono selezionate le sequenze con mascherina come training set e le sequenze senza mascherina come test set.
(3) **Tipo 3:** Vengono selezionate casualmente, per ogni soggetto, il 70% delle sequenze come training set e le restanti 30% come test set. Quindi entrambi i set contengono sia sequenze con mascherina che senza.

Per quanto riguarda i dataset, vengono utilizzati come riferimento:

(1) **M2FRED:** Già descritto in precedenza ed i cui dati sono riassunti in Tabella 1.
(2) **XM2VTS:** Anch'esso precedentemente presentato ed i cui dati sono mostrati in Tabella 2.
(3) **M2FRED Ridotto:** Viene applicata una riduzione orizzontale sui dati di M2FRED, diminuendo così il numero di sequenze (sia con mascherina che senza) da 15 a 5 per ogni soggetto, in modo tale da adattarle ad XM2VTS. In Tabella 6 sono riportati in modo sintetico i dati associati al dataset.
(4) **XM2VTS Ridotto:** Viene attuata una riduzione verticale sui dati di XM2VTS, diminuendo così il numero di soggetti totali da 294 a 46, in modo da adattarli ad M2FRED. In particolare, i soggetti sono scelti in modo tale da uguagliare il numero di soggetti di sesso maschile e femminile di M2FRED, così che la rete utilizzata possa apprendere le caratteristiche del volto maschile e femminile nello stesso modo di M2FRED ridotto. In Tabella 6 sono riportati sinteticamente i dati associati al dataset.

Le 3 tipologie di esperimenti vengono quindi applicate a tutti e 4 i dataset di riferimento, utilizzando come architettura il Modello 2 con i layers della rete VGG16 freezed ed i seguenti parametri:

- **No. unità LSTM:** 512
- **batch size:** 4
- **learning rate:** 0.0001
- **epoche:** 325

Si nota quindi che gli esperimenti di Tipo 1 e Tipo 2 utilizzano sempre training set e test set in rapporto 50:50; mentre gli esperimenti di Tipo 3 lavorano su training set e test set in rapporto 70:30.

Num. Soggetti	Seq. con mascherina	Seq. senza mascherina	Num. Frames	Dim. Frame
46	5	5	50	200x200 px

Table 6: M2FRED ed XM2VTS Ridotti

4.4.1 Risultati relativi ad M2FRED.

	precision	recall	f1-score	accuracy
macro avg	0.91	0.88	0.88	
weighted avg	0.91	0.88	0.88	
				0.88

Table 7: M2FRED Esperimento Tipo 1: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.91	0.91	0.90	
weighted avg	0.91	0.91	0.90	
				0.91

Table 8: M2FRED Esperimento Tipo 2: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	1.00	1.00	1.00	
weighted avg	1.00	1.00	1.00	
				1.00

Table 9: M2FRED Esperimento Tipo 3: accuracy, precision, recall ed F1 score.

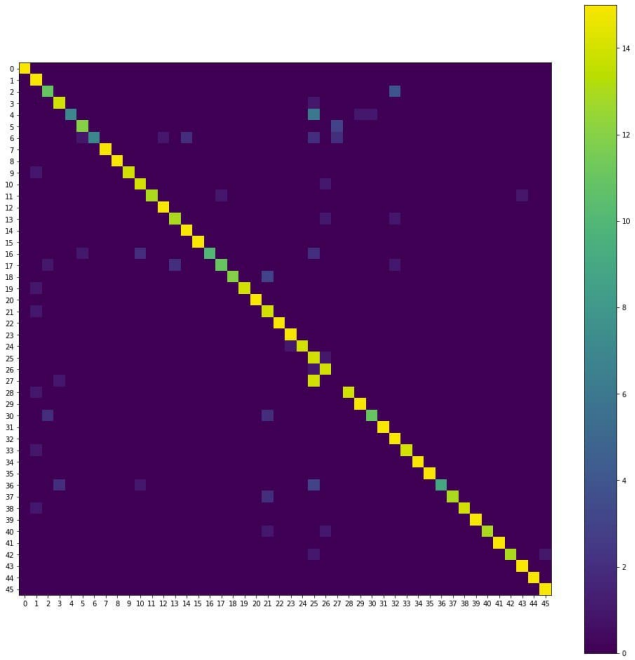

Figure 13: M2FRED Esperimento Tipo 1: Confusion Matrix

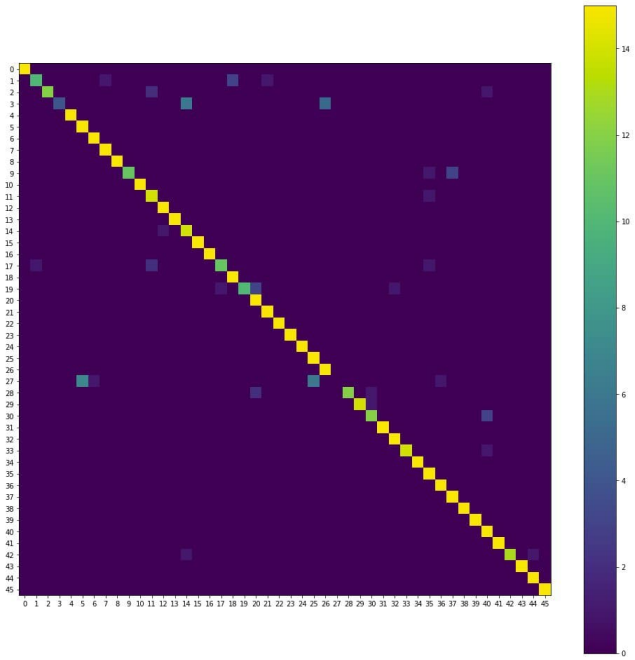

Figure 14: M2FRED Esperimento Tipo 2: Confusion Matrix

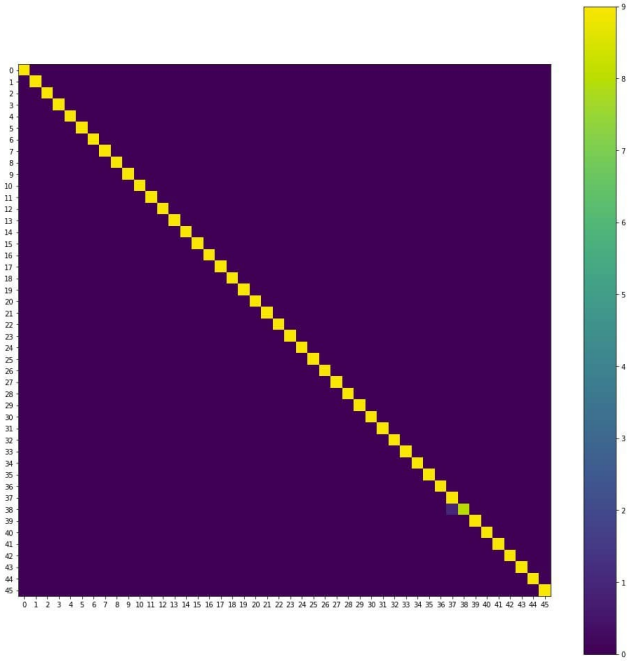

Figure 15: M2FRED Esperimento Tipo 3: Confusion Matrix

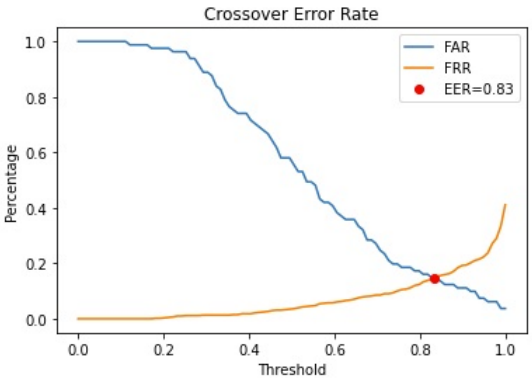

Figure 16: M2FRED Esperimento Tipo 1: FAR, FRR, EER

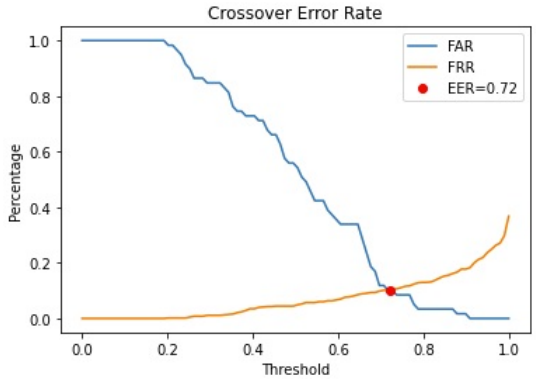

Figure 17: M2FRED Esperimento Tipo 2: FAR, FRR, EER

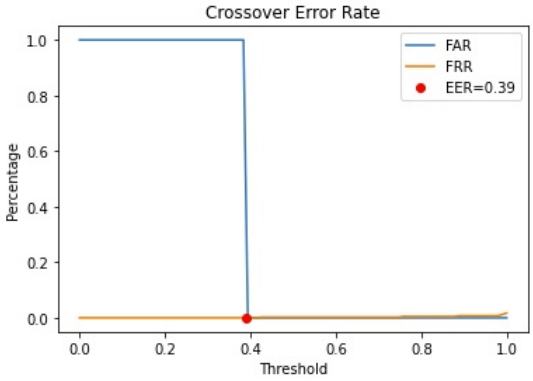

Figure 18: M2FRED Esperimento Tipo 3: FAR, FRR, EER

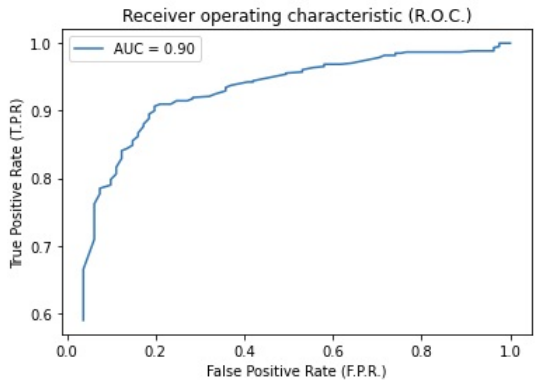

Figure 19: M2FRED Esperimento Tipo 1: ROC Curve

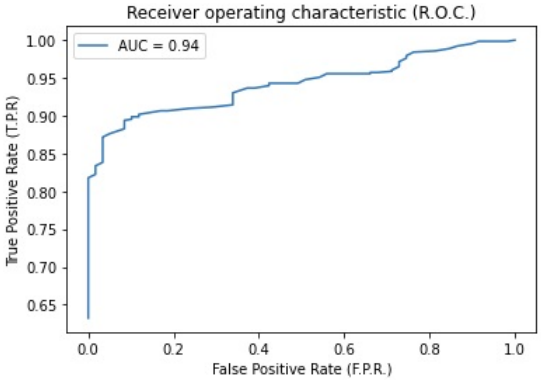

Figure 20: M2FRED Esperimento Tipo 2: ROC Curve

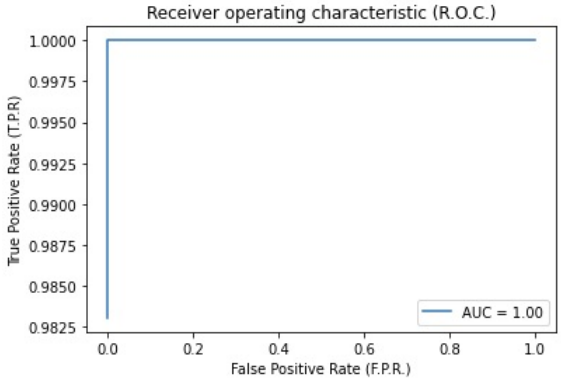

Figure 21: M2FRED Esperimento Tipo 3: ROC Curve

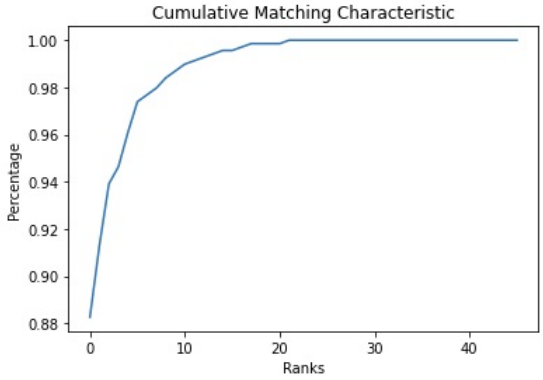

Figure 22: M2FRED Esperimento Tipo 1: CMC Curve

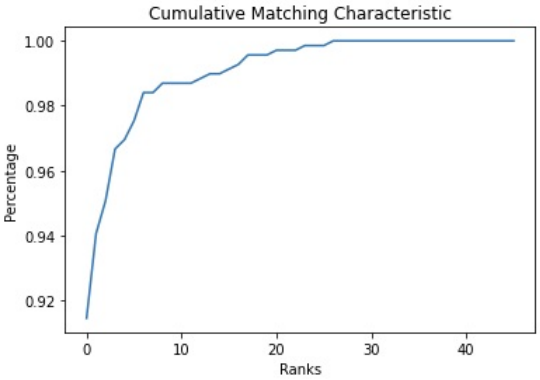

Figure 23: M2FRED Esperimento Tipo 2: CMC Curve

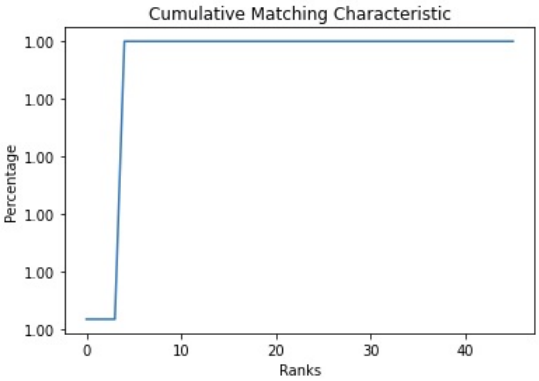

Figure 24: M2FRED Esperimento Tipo 3: CMC Curve

4.4.2 Risultati relativi ad M2FRED Ridotto.

	precision	recall	f1-score	accuracy
macro avg	0.85	0.81	0.80	
weighted avg	0.85	0.81	0.80	
				0.81

Table 10: M2FRED Ridotto Esperimento Tipo 1: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.83	0.82	0.80	
weighted avg	0.83	0.82	0.80	
				0.82

Table 11: M2FRED Ridotto Esperimento Tipo 2: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.96	0.93	0.93	
weighted avg	0.96	0.93	0.93	
				0.93

Table 12: M2FRED Ridotto Esperimento Tipo 3: accuracy, precision, recall ed F1 score.

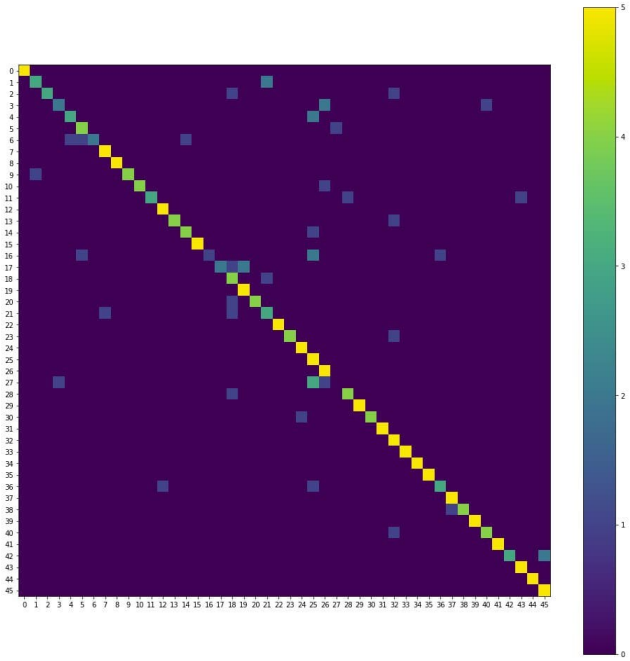

Figure 25: M2FRED Ridotto Esperimento Tipo 1: Confusion Matrix

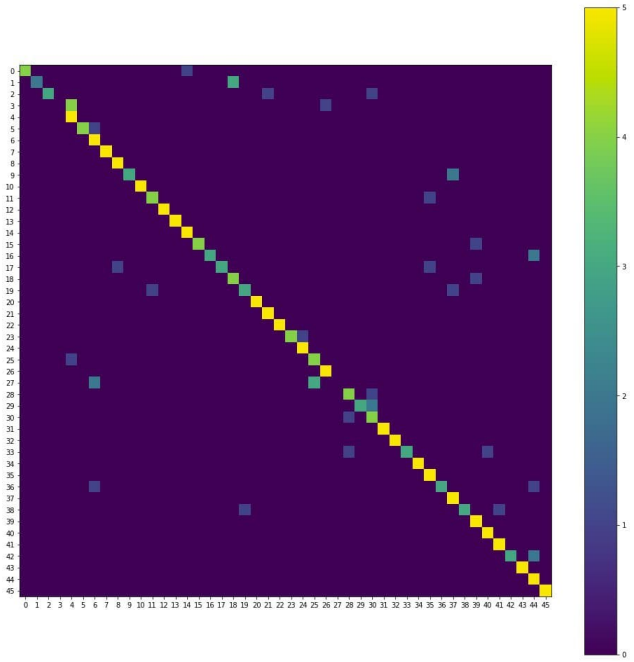

Figure 26: M2FRED Ridotto Esperimento Tipo 2: Confusion Matrix

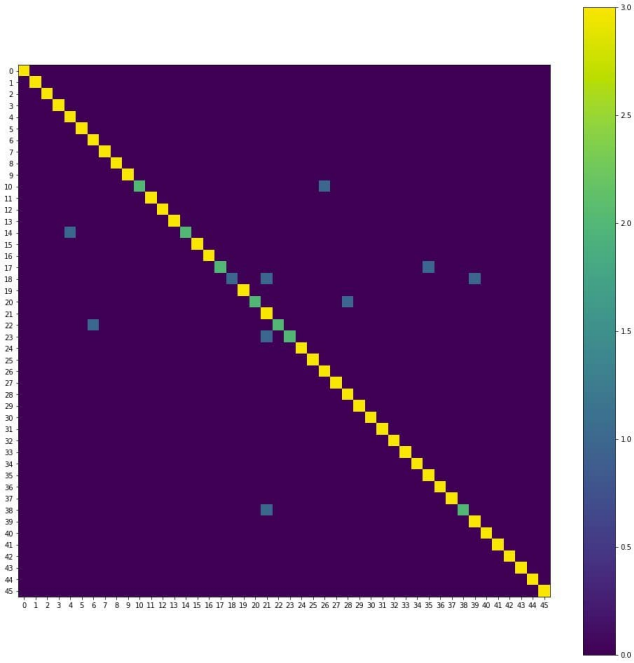

Figure 27: M2FRED Ridotto Esperimento Tipo 3: Confusion Matrix

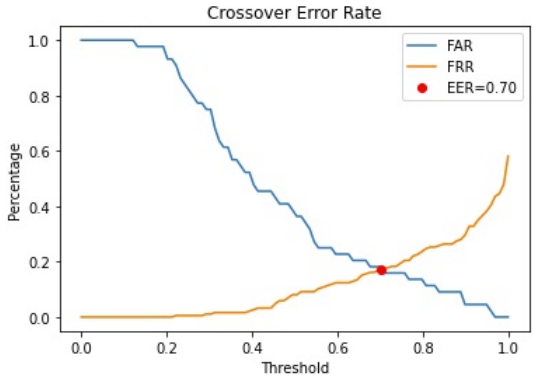

Figure 28: M2FRED Ridotto Esperimento Tipo 1: FAR, FRR, EER

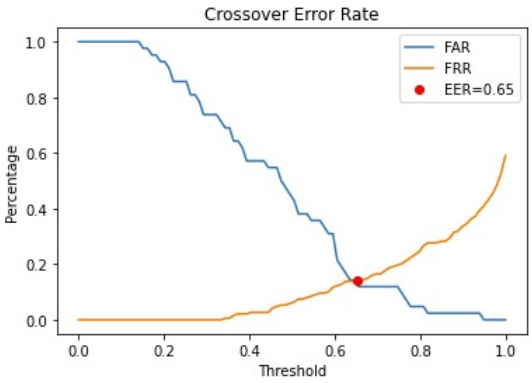

Figure 29: M2FRED Ridotto Esperimento Tipo 2: FAR, FRR, EER

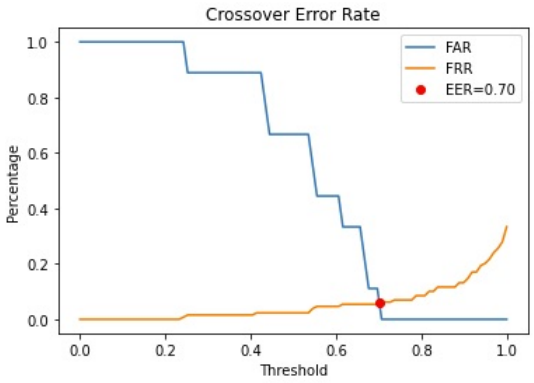

Figure 30: M2FRED Ridotto Esperimento Tipo 3: FAR, FRR, EER

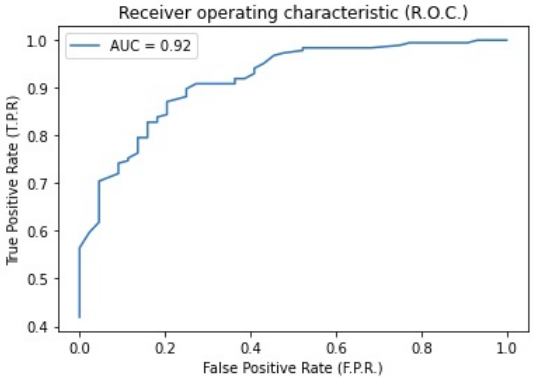

Figure 31: M2FRED Ridotto Esperimento Tipo 1: ROC Curve

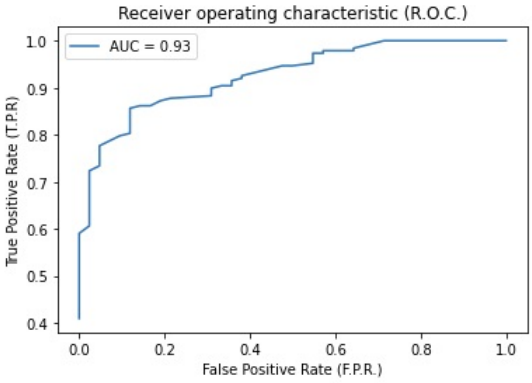

Figure 32: M2FRED Ridotto Esperimento Tipo 2: ROC Curve

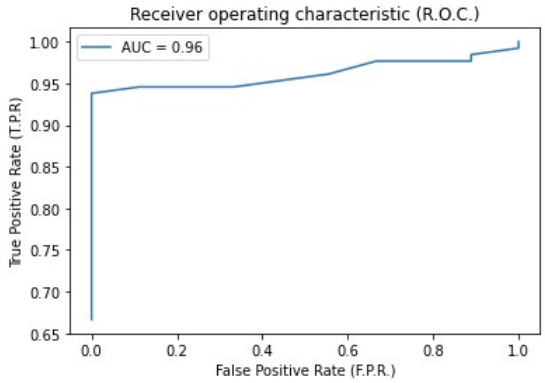

Figure 33: M2FRED Ridotto Esperimento Tipo 3: ROC Curve

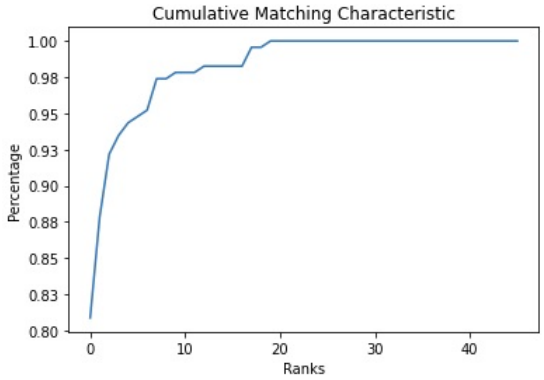

Figure 34: M2FRED Ridotto Esperimento Tipo 1: CMC Curve

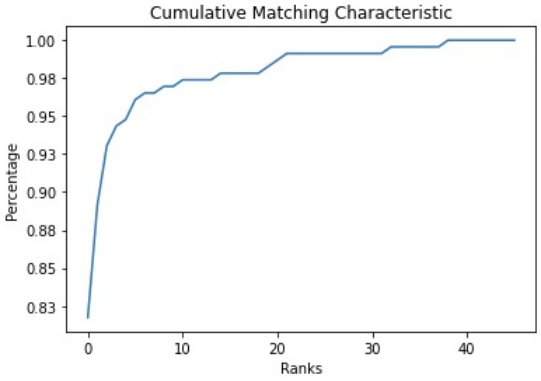

Figure 35: M2FRED Ridotto Esperimento Tipo 2: CMC Curve

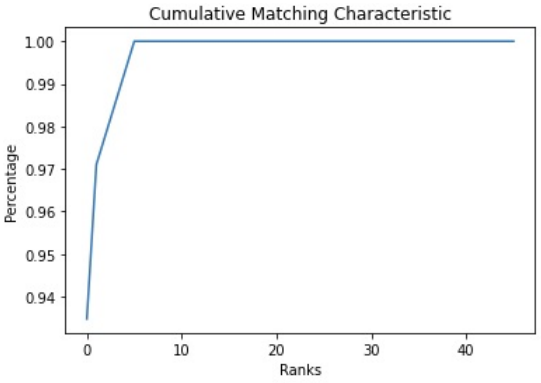

Figure 36: M2FRED Ridotto Esperimento Tipo 3: CMC Curve

4.4.3 Risultati relativi ad XM2VTS.

	precision	recall	f1-score	accuracy
macro avg	0.92	0.91	0.91	
weighted avg	0.92	0.91	0.91	
				0.91

Table 13: XM2VTS Esperimento Tipo 1: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.91	0.91	0.90	
weighted avg	0.91	0.91	0.90	
				0.91

Table 14: XM2VTS Esperimento Tipo 2: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	1.00	1.00	1.00	
weighted avg	1.00	1.00	1.00	
				1.00

Table 15: XM2VTS Esperimento Tipo 3: accuracy, precision, recall ed F1 score.

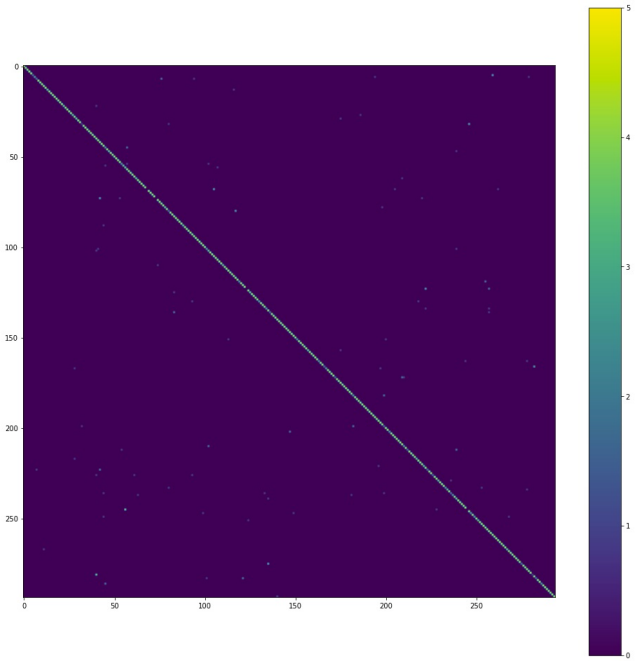

Figure 37: XM2VTS Esperimento Tipo 1: Confusion Matrix

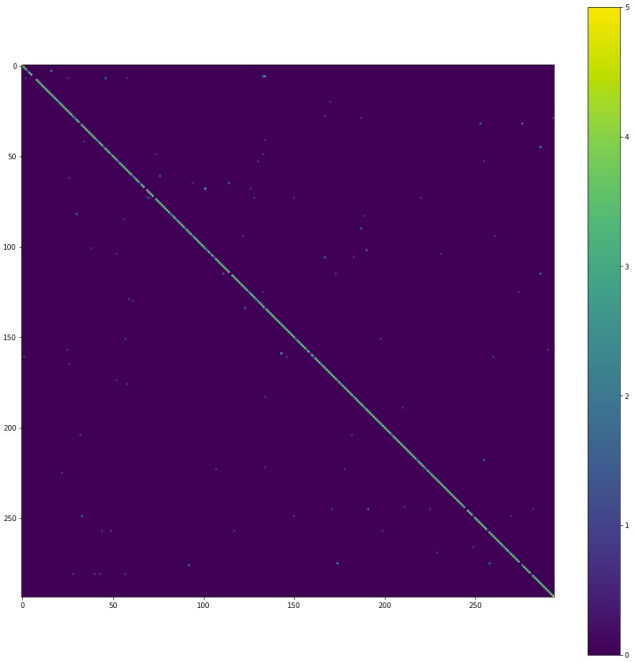

Figure 38: XM2VTS Esperimento Tipo 2: Confusion Matrix

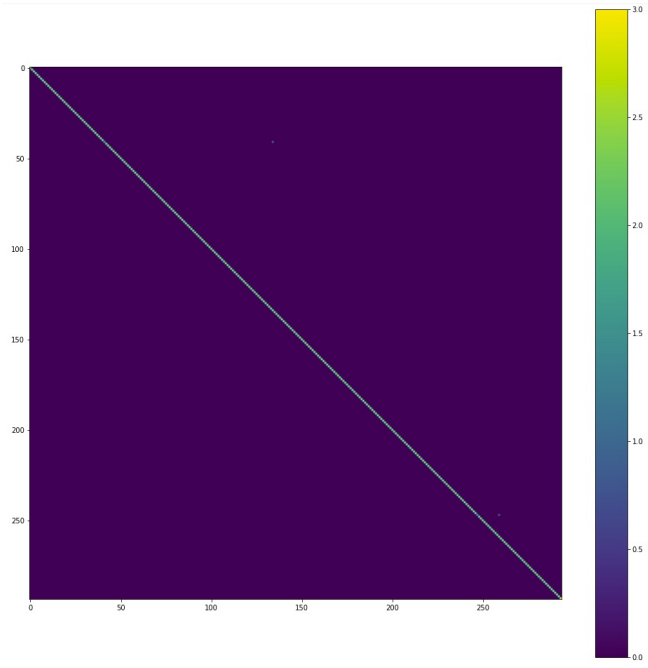

Figure 39: XM2VTS Esperimento Tipo 3: Confusion Matrix

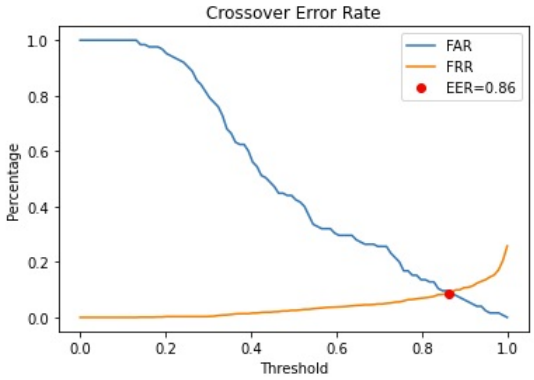

Figure 40: XM2VTS Esperimento Tipo 1: FAR, FRR, EER

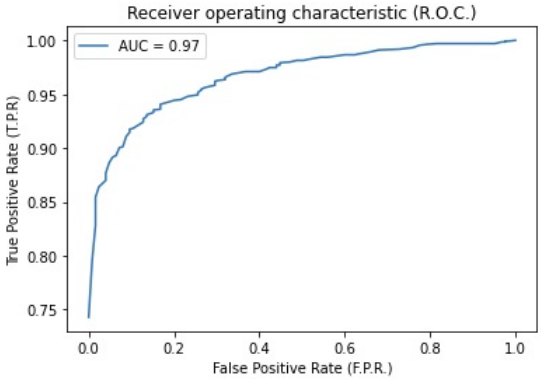

Figure 43: XM2VTS Esperimento Tipo 1: ROC Curve

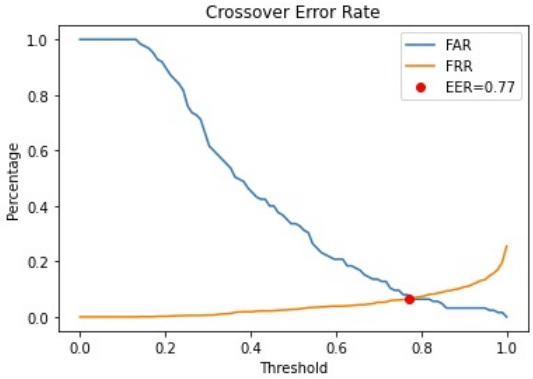

Figure 41: XM2VTS Esperimento Tipo 2: FAR, FRR, EER

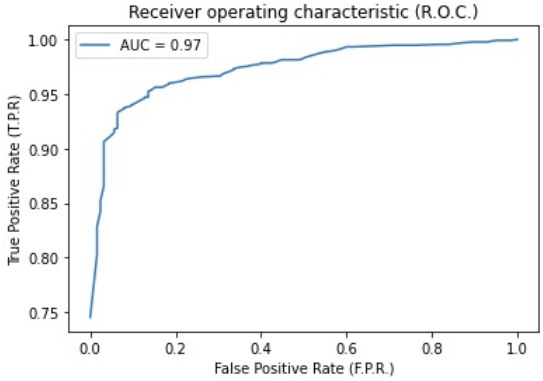

Figure 44: XM2VTS Esperimento Tipo 2: ROC Curve

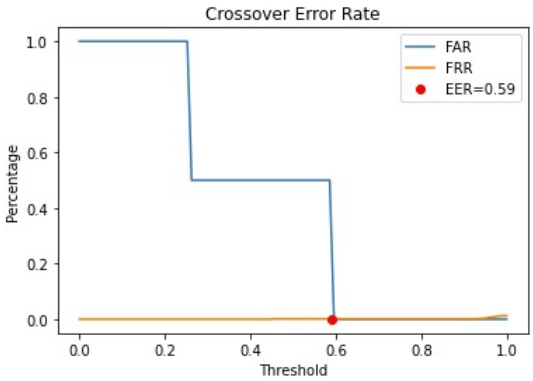

Figure 42: XM2VTS Esperimento Tipo 3: FAR, FRR, EER

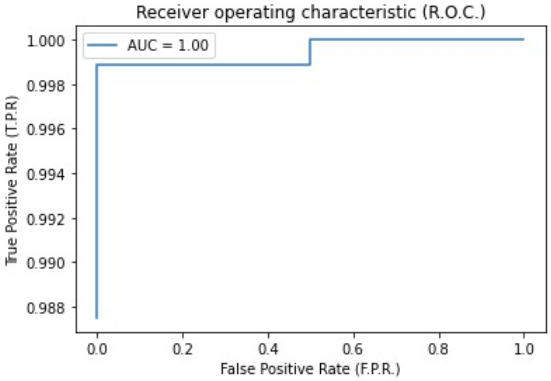

Figure 45: XM2VTS Esperimento Tipo 3: ROC Curve

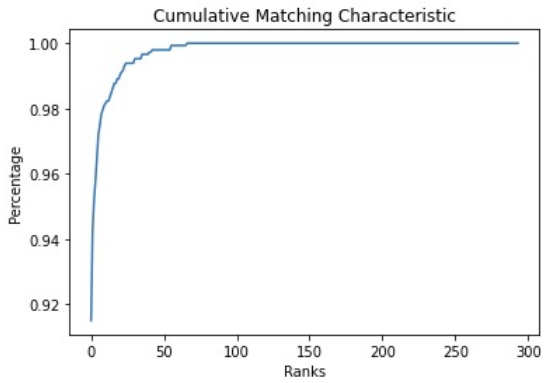

Figure 46: XM2VTS Esperimento Tipo 1: CMC Curve

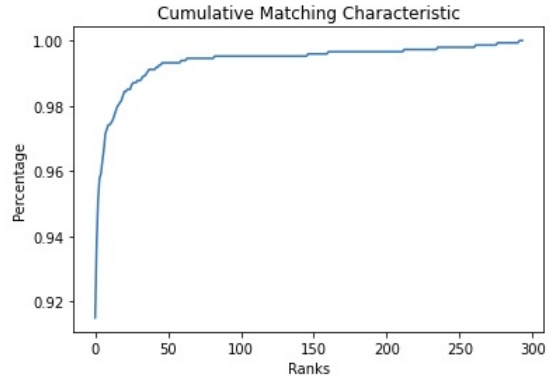

Figure 47: XM2VTS Esperimento Tipo 2: CMC Curve

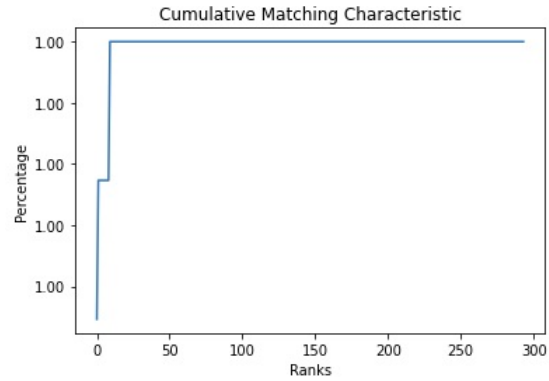

Figure 48: XM2VTS Esperimento Tipo 3: CMC Curve

4.4.4 Risultati relativi ad XM2VTS Ridotto.

	precision	recall	f1-score	accuracy
macro avg	0.98	0.97	0.97	
weighted avg	0.98	0.97	0.97	
				0.97

Table 16: XM2VTS Ridotto Esperimento Tipo 1: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.97	0.96	0.95	
weighted avg	0.97	0.96	0.95	
				0.96

Table 17: XM2VTS Ridotto Esperimento Tipo 2: accuracy, precision, recall ed F1 score.

	precision	recall	f1-score	accuracy
macro avg	0.99	0.99	0.99	
weighted avg	0.99	0.99	0.99	
				0.99

Table 18: XM2VTS Ridotto Esperimento Tipo 3: accuracy, precision, recall ed F1 score.

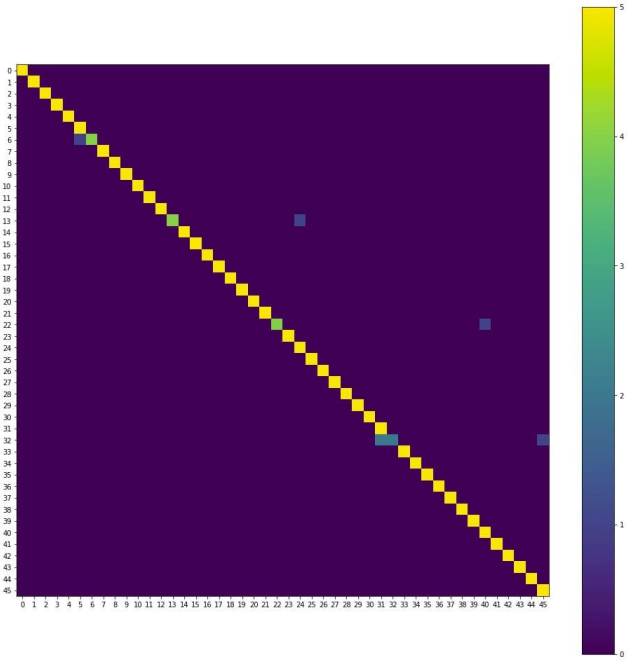

Figure 49: XM2VTS Ridotto Esperimento Tipo 1: Confusion Matrix

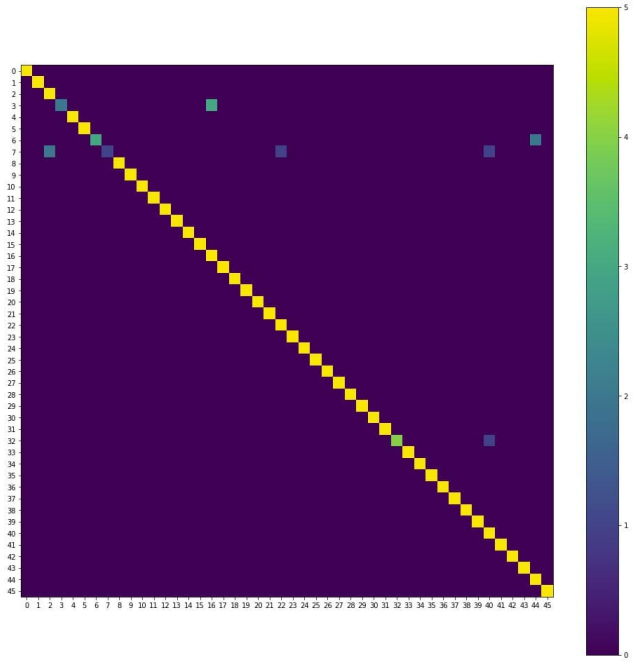

Figure 50: XM2VTS Ridotto Esperimento Tipo 2: Confusion Matrix

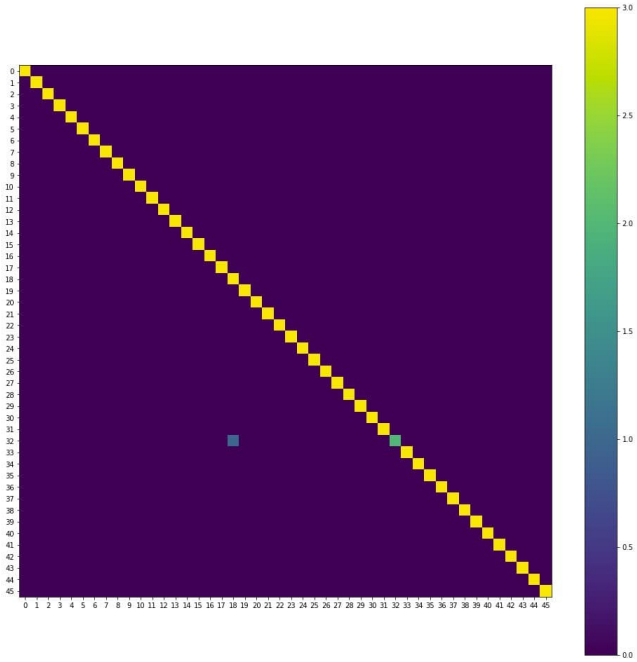

Figure 51: XM2VTS Ridotto Esperimento Tipo 3: Confusion Matrix

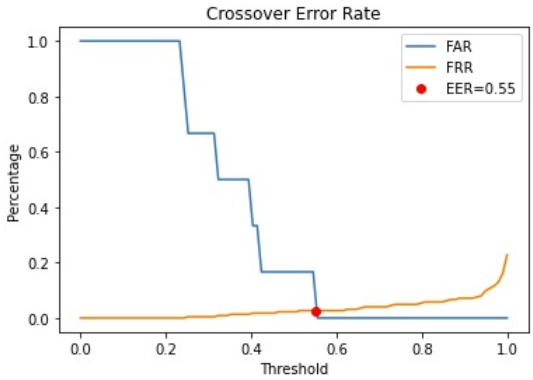

Figure 52: XM2VTS Ridotto Esperimento Tipo 1: FAR, FRR, EER

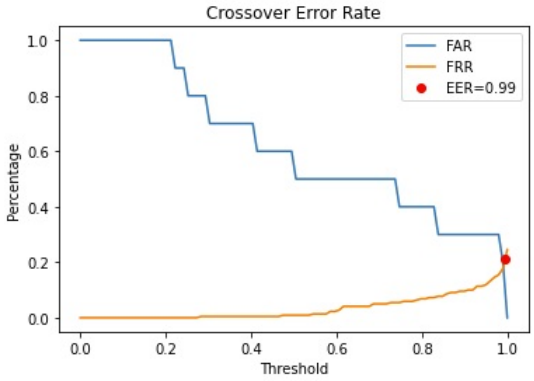

Figure 53: XM2VTS Ridotto Esperimento Tipo 2: FAR, FRR, EER

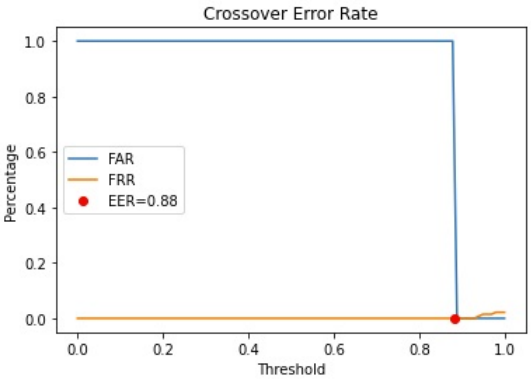

Figure 54: XM2VTS Ridotto Esperimento Tipo 3: FAR, FRR, EER

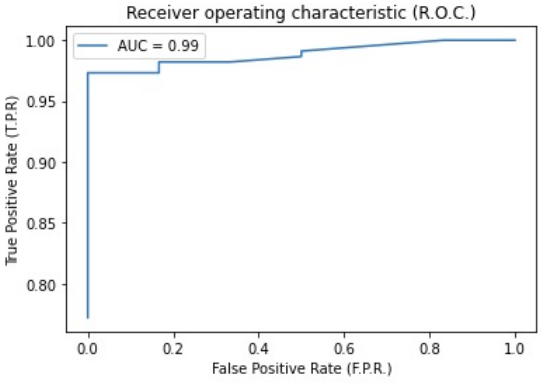

Figure 55: XM2VTS Ridotto Esperimento Tipo 1: ROC Curve

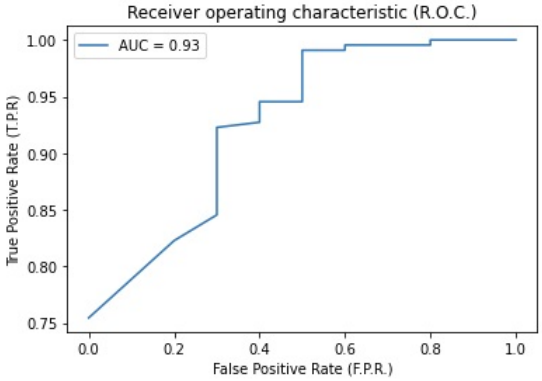

Figure 56: XM2VTS Ridotto Esperimento Tipo 2: ROC Curve

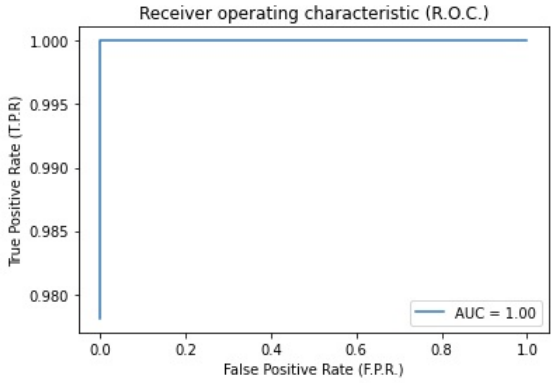

Figure 57: XM2VTS Ridotto Esperimento Tipo 3: ROC Curve

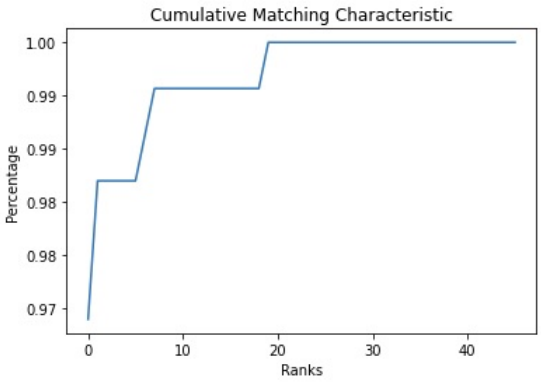

Figure 58: XM2VTS Ridotto Esperimento Tipo 1: CMC Curve

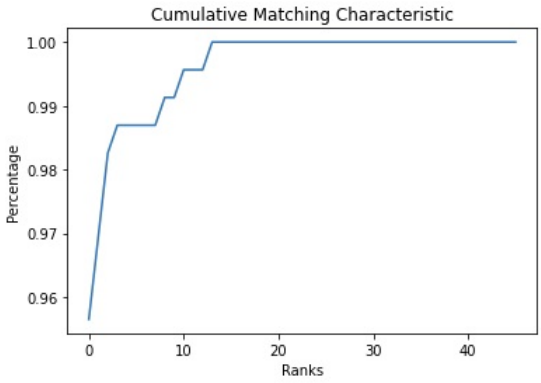

Figure 59: XM2VTS Ridotto Esperimento Tipo 2: CMC Curve

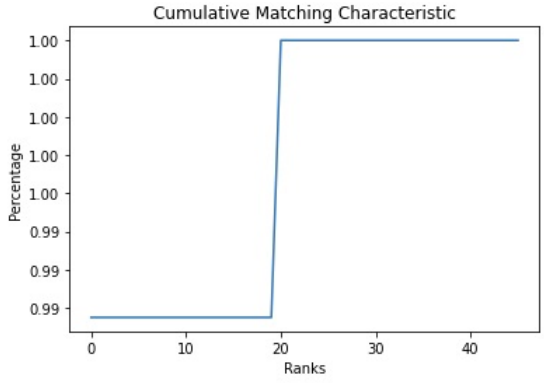

Figure 60: XM2VTS Ridotto Esperimento Tipo 3: CMC Curve

5 DISCUSSIONE DEI RISULTATI

Sulla base degli esperimenti effettuati, si procede dunque ad un'analisi dei risultati ottenuti.

5.1 Modello 1 vs Modello 2

Sulla base dei grafici riportati in Figura 10 ed in Figura 11 e dei risultati in Tabella 3, si apprende che il Modello 2 produce risultati leggermente migliori rispetto al Modello 1. La causa di questo incremento è riconducibile al fatto che la struttura VGG16 nel secondo modello è specificatamente addestrata nel riconoscere i volti, dal momento che il dataset VGGFace contiene esclusivamente i suddetti; a differenza della VGG16 presente nel primo modello, addestrata su un dataset più variegato.

5.2 Tuning degli iperparametri

Dai risultati presentati in Tabella 4, si osserva che il valore di learning rate più elevato produce i risultati peggiori, rispetto a quello più basso ed a quello intermedio, per il quale sono stati invece ottenuti i migliori risultati. Possiamo osservare inoltre che l'aumento del numero di unità LSTM conduce generalmente ad un incremento significativo dell'accuratezza fino al valore 256, dopo il quale si registrano incrementi poco significativi o addirittura dei decrementi. Sulla base di questi risultati e del grafico in Figura 11, osserviamo che il massimo valore di accuratezza è raggiunto con la seguente configurazione:

- **No. unità LSTM:** 512
- **batch size:** 4
- **learning rate:** 0.0001
- **epoche:** 325

5.3 Unfreezing dei layers

Dai risultati presentati in Figura 12 ed in Figura 11 e dalla Tabella 5, osserviamo che il Modello con i layer unfreezed produce risultati drasticamente inferiori. La causa è riconducibile quindi al fatto che la rete VGG16 non riesce ad apprendere le caratteristiche spaziali sui dati di riferimento, a causa della esigua quantità di questi ultimi e della qualità degli stessi.

5.4 Sperimentazioni relative ai dataset

Si procede ora all'analisi dei dati relativi alle sperimentazioni eseguite sui dataset.

5.4.1 M2FRED.

In relazione ai risultati mostrati in Tabella 7 e Tabella 8 ed ai grafici in Figura 13, 22, 16, 19, 14, 23, 17, 20, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 2 non sembrano esserci differenze significative in termini prestazionali. Dunque sembrerebbe che l'utilizzo di sequenze con mascherina o senza mascherina nel processo di training sia efficace in ugual misura. Questo risultato può essere giustificato in virtù del fatto che la soluzione adottata elimina quasi completamente la mascherina dal processo di estrazione delle caratteristiche, dal momento che viene considerata la sola regione perioculare.

Per quanto concerne i risultati mostrati in Tabella 7 e Tabella 9 ed i grafici in Figura 13, 22, 16, 19, 15, 24, 18, 21, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 3 sembra esserci un aumento significativo dei parametri prestazionali, raggiungendo valori del 100%. Da ciò si ipotizza dunque che l'aumento dei dati di training ed il conseguente decremento di quelli di testing, abbiano condotto la rete in overfit.

5.4.2 M2FRED Ridotto.

In relazione ai risultati mostrati in Tabella 10 e Tabella 11 ed ai grafici in Figura 25, 34, 28, 31, 26, 35, 29, 32, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 2 non sembrano esserci differenze significative in termini prestazionali, rafforzando quindi l'ipotesi precedentemente esposta.

Per quanto concerne i risultati mostrati in Tabella 10 e Tabella 12 ed i grafici in Figura 25, 34, 28, 31, 27, 36, 30, 33, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 3 sembra esserci un aumento significativo dei parametri prestazionali, riconducibile all'aumento dei dati di training. Ad ogni modo, non è comunque da escludere la possibilità di overfitting da parte della rete.

5.4.3 XM2VTS.

In relazione ai risultati mostrati in Tabella 13 e Tabella 14 ed ai grafici in Figura 37, 46, 40, 43, 38, 47, 41, 44, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 2 non sembrano esserci differenze significative in termini prestazionali, corroborando ulteriormente l'ipotesi formulata.

Per quanto concerne i risultati mostrati in Tabella 13 e Tabella 15 ed i grafici in Figura 37, 46, 40, 43, 39, 48, 42, 45, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 3 sembra esserci un aumento significativo dei parametri prestazionali, raggiungendo valori del 100%. Per le stesse ragioni precedenti è possibile quindi avanzare l'ipotesi di overfit.

5.4.4 XM2VTS Ridotto.

In relazione ai risultati mostrati in Tabella 16 e Tabella 17 ed ai grafici in Figura 49, 58, 52, 55, 50, 59, 53, 56, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 2 non sembrano esserci differenze significative in termini prestazionali, rafforzando ancora una volta l'ipotesi enunciata in precedenza.

Per quanto concerne i risultati mostrati in Tabella 16 e Tabella 18 ed i grafici in Figura 49, 58, 52, 55, 51, 60, 54, 57, si osserva che tra le sperimentazioni di Tipo 1 e Tipo 3 sembra esserci un aumento significativo dei parametri prestazionali, sfiorando i valori del 100%. Per le stesse ragioni precedenti è possibile quindi avanzare l'ipotesi di overfit.

5.4.5 M2FRED vs M2FRED Ridotto.

In relazione ai risultati mostrati in sezione 4.4.1 ed in sezione 4.4.2, si osserva che la riduzione del numero di sequenze per ogni soggetto ha condotto ad una riduzione dei parametri prestazionali. Il fenomeno è del tutto regolare, dal momento che le prestazioni di una rete incrementano assieme alla quantità di dati (viceversa nel caso del decremento dei dati).

5.4.6 *XM2VTS vs XM2VTS Ridotto.*

In relazione ai risultati mostrati in sezione 4.4.3 ed in sezione 4.4.4, si osserva che la riduzione del numero di soggetti ha condotto ad un aumento dei parametri prestazionali. La causa potrebbe essere riconducibile al fatto che avendo meno soggetti, è più difficile per la rete confondere il volto di un soggetto con un altro, producendo quindi meno errori.

5.4.7 *M2FRED Ridotto vs XM2VTS Ridotto.*

Il relazione ai risultati mostrati in sezione 4.4.2 ed in sezione 4.4.4, si osserva che XM2VTS Ridotto ottiene prestazioni migliori rispetto ad M2FRED Ridotto. Questo comportamento potrebbe essere riconducibile al fatto che le sequenze in XM2VTS sono state acquisite in un ambiente controllato, presentando quindi poca variabilità in termini di illuminazione e posa, e quindi la rete riesce ad estrarre le caratteristiche con maggiore facilità.

6 CONCLUSIONI

In conclusione, sembra che la soluzione proposta permetterebbe di risolvere il problema del Masked Face Recognition, dal momento che i risultati ottenuti sembrano essere soddisfacenti e, inoltre, ci consentirebbe di trattare indistintamente sequenze con mascherina e sequenze senza mascherina. D'altro canto l'architettura utilizzata sembrerebbe soffrire di overfitting nel caso in cui vengano utilizzati più dati di training che di testing.

REFERENCES

[1] Asit Kumar Datta, Madhura Datta, and Pradipta Kumar Banerjee. 2015. *Face Detection and Recognition: Theory and Practice*. (1st. edition). CRC.

[2] G. N. Priya and R. W. Banu. 2014. Occlusion invariant face recognition using mean based weight matrix and support vector machine, 303–315.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

[4] S. Almabdy and L. Elrefaei. 2019. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*.

[5] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu. 2019. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. *Proceedings of the IEEE International Conference on Computer Vision*, 773–782.

[6] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. 2010. Action classification in soccer videos with long short-term memory recurrent neural networks. *Artificial Neural Networks-ICANN*, 154–159.

[7] Prodesire. 2020. Face-mask. https://github.com/Prodesire/face-mask.

[8] dangz90. 2018. Deep learning for expression recognition. https://github.com/dangz90/Deep-Learning-for-Expression-Recognition-in-Image-Sequences.