



PROGETTO FONDAMENTI DI VISIONE ARTIFICIALE E BIOMETRIA

**DeepFake Detection attraverso l'utilizzo del
modello CLIP**

Autori

Chiara Puglia 0522501984

Luca Giuliano 0522501931

Antonio Landi 0522502075

Università degli Studi di Salerno

a.a 2024/2025

1 Stato dell’arte

Negli ultimi anni si sono moltiplicati gli approcci di apprendimento automatico per distinguere immagini reali da Deepfake. Alcuni lavori sfruttano reti Siamesi con triplet loss, che apprendono a mappare immagini reali vicine nello spazio delle caratteristiche e quelle fake distanti. Ad esempio, Kanwal et al. (2023) propongono una rete Siamese con triplet loss allenata su immagini GAN, in cui ogni tripla (anchor, positive, negative) guida l’apprendimento delle caratteristiche distintive tra volti reali e sintetici. I loro esperimenti mostrano che questa architettura triplet supera nettamente metodi contemporanei, ottenendo oltre il 90% di accuratezza nella classificazione tra immagini reali e Deepfake.

In modo simile, Liang et al. (2023) integrano mappe di profondità predette nel processo di estrazione delle caratteristiche: il loro modello unisce un modulo di stima della profondità (che cattura discontinuità tipiche dei volti Deepfake) con un estrattore di feature soggetto a triplet loss. In questo schema, coppie di volti reali appaiono vicine nello spazio latente mentre coppie reali-fake rimangono lontane. I risultati su dataset pubblici (FaceForensics++ e Celeb-DF) confermano la superiorità di questo approccio depth-guided rispetto a metodi di confronto standard, validando l’efficacia del triplet loss nel separare reali e falsi.

Negli ultimi anni sono emersi anche metodi multimodali basati su grandi modelli visione-linguaggio. In particolare, Ojha et al. (2023) (CVPR 2023) hanno sfruttato il modello CLIP pre-addestrato per ottenere features visive generali: congelano l’encoder di CLIP e ne addestrano solo un classificatore lineare finale. Questo approccio “linear probing” su CLIP fornisce elevata capacità di generalizzazione: i risultati ottenuti da Ojha et al. con CLIP+linear probe mostrano prestazioni allo stato dell’arte sia su immagini GAN-generated sia su immagini generate da diffusion model. Il punto di forza è che CLIP, pur non essendo stato specificamente addestrato per riconoscere Deepfake, conserva robusti pattern discriminativi per immagini sintetiche.

In sintesi, mentre alcuni lavori recenti impiegano modelli multimodali (CLIP) per catturare informazioni globali e linguistiche, la maggior parte delle soluzioni di spicco nel periodo 2022–2025 combina estrattori di feature specializzati, riduzione di dimensionalità (es. PCA) e classificatori SVM, oppure architetture con triplet loss per separare efficacemente immagini reali da Deepfake.

1.1 Dataset

Il dataset è costituito da 2496738 immagini di cui 964989 rappresentano le immagini reali mentre 1531749 rappresentano le immagini fake. Per la generazione delle immagini fake, sono stati utilizzati 25 generatori (come 13 GANs, 7 Diffusion, e 5 miscellaneous generators) e per quanto riguarda le categorie incluse nel dataset sono: Human/Human Faces, Animal/Animal Faces, Places, Vehicles, Art, e altri oggetti real-life. Il dataset, inoltre, è stato suddiviso in 32 cartelle, ognuna delle quali contiene un file metadata.csv all’interno del quale sono state definite le seguenti informazioni:

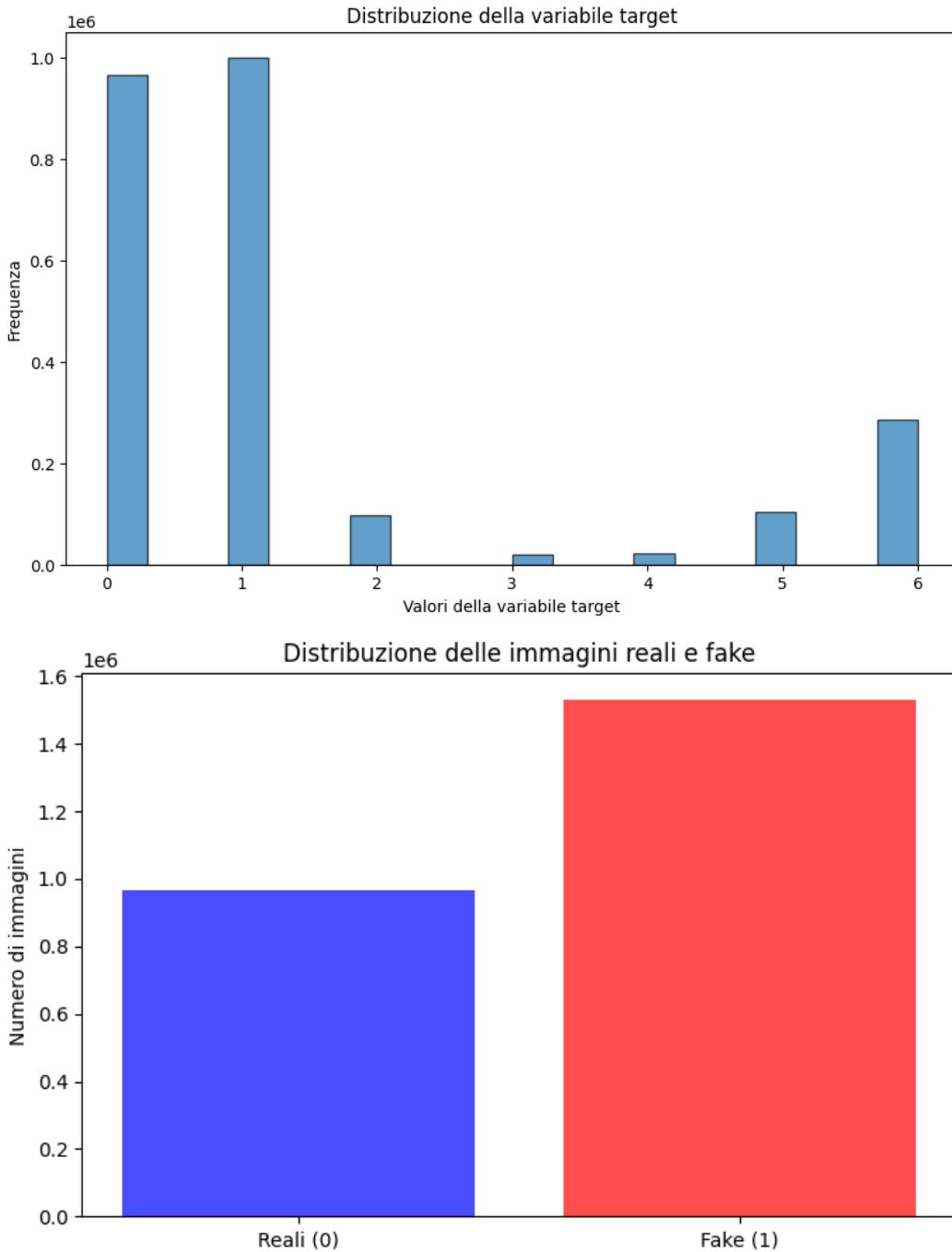
- 1. filename:** Rappresenta il nome dell'immagine.
- 2. image_path:** Rappresenta il percorso fino alla cartella in cui è situata l'immagine.
- 3. target:** Rappresentato da un numero che va da 0 a 6 dove 0 indica le immagini reali mentre tutti gli altri valori maggiori di 0 rappresentano le immagini fake.
- 4. category:** Rappresentata da una stringa, è associata all'immagine di riferimento e ne rappresenta il tipo.

Le immagini vengono campionate casualmente utilizzando diversi metodi e poi trasformate tramite compromessi. Tutte le immagini sono passate attraverso RandomCrop e Random Impairments (compressione Jpeg e ridimensionamento). Per applicare queste trasformazioni, è stato utilizzato transform.py, che applica trasformazioni casuali. Tutte le immagini sono ritagliate e ridimensionate a 200×200 pixel e poi compresse usando JPEG a un livello di qualità casuale.

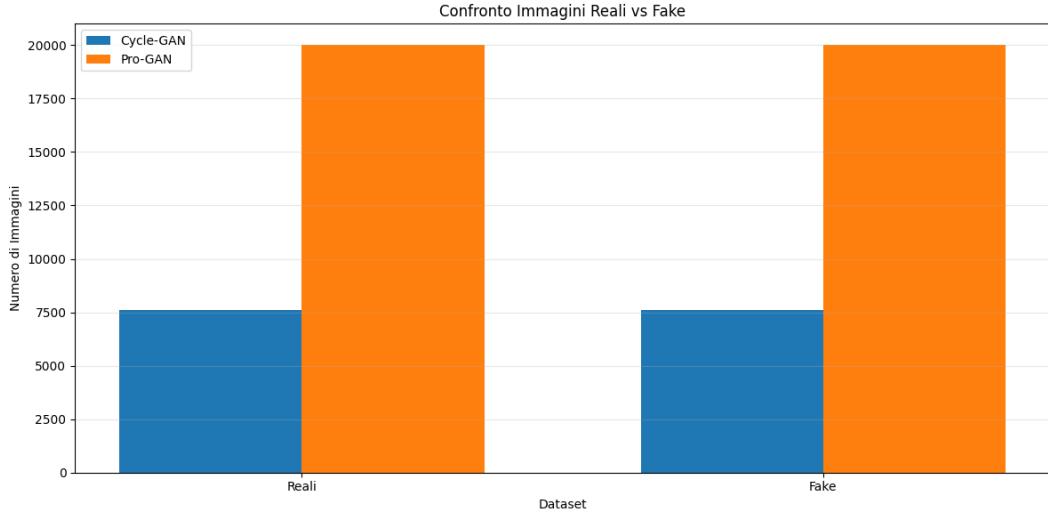
2 Pre-Processing: Scelta dei dataset

Sulla base delle informazioni esposte nello stato dell'arte, l'obiettivo di questa ricerca consiste nell'addestrare un modello che deve essere in grado di riconoscere se un'immagine è real o fake. Data la premessa, in questa sezione verrà affrontata tutta la fase di pre-processing a cui seguirà quelle che sono state le scelte progettuali. In relazione a ciò, sono state effettuate le seguenti analisi:

Per quanto riguarda la feature target, in seguito ad un'analisi esplorativa dei dataset, tutti i valori maggiori di 0, ovvero da 1 a 6 per questa feature, sono stati convertiti in 1 dal momento che tutte le immagini che presentavano target maggiore di 0 erano immagini fake. A tal proposito, è stato riportato il grafico delle distribuzioni dei valori assunti dalla variabile target nei dataset e in seguito alle conversioni effettuate, il grafico con il numero totale delle immagini reali e fake presenti nei dataset:



La maggior parte dei dataset presentano un numero di immagini reali e fake sbilanciato, ma di questi dataset, solo 2 in particolare presentano bilanciamento tra immagini reali ed immagini fake e sono rispettivamente Cycle-Gan e Pro-Gan. Entrambi i dataset sono stati utilizzati per l'analisi. In seguito, è stato riportato un grafico che rappresenta il numero totale di immagini reali e fake che costituiscono sia il dataset Cycle-Gan che il dataset Pro-Gan:



Come si può evincere dal grafico, il dataset Cycle-Gan presenta 7605 immagini reali e 7605 immagini fake, mentre il dataset Pro-Gan presenta 20000 immagini reali e altrettanti 20000 immagini fake, risultando così bilanciati.

3 Scelte progettuali

3.1 Riconoscimento di Immagini Real e Fake tramite l'estrazione delle informazioni per le sole immagini

In questa sezione, ci concentreremo principalmente su come sono state acquisite le immagini utilizzando il modello CLIP (Contrastive Language-Image Pretraining) che, a sua volta, carica un transformer Vit-B/32 e come quest'ultimo è stato addestrato per farsi che tale modello sia in grado di fare distinzione tra immagini reali e fake. In seguito, è stato riportato un grafico che introduce brevemente quelle che sono state le fasi di progettazione con conseguente focus su ognuna di esse.



La prima fase consiste nella conversione delle immagini in vettori di embeddings di dimensione 512, che successivamente sono stati ridotti a 128 per proiettare l'embedding dell'immagine in uno spazio di dimensione ridotta. Tale conversione è stata effettuata dal momento che, tutti i modelli compreso anche CLIP, non sono in grado di lavorare direttamente con le immagini. Prima dell'estrazione dei vettori di embeddings dalle immagini, esse vengono denormalizzate, poichè i modelli come CLIP normalizzano le immagini usando valori specifici di media e deviazione standard (es: mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). Questa normalizzazione trasforma i pixel in un range non visibile direttamente. Senza denormalizzazione, le immagini apparirebbero distorte o completamente nere.

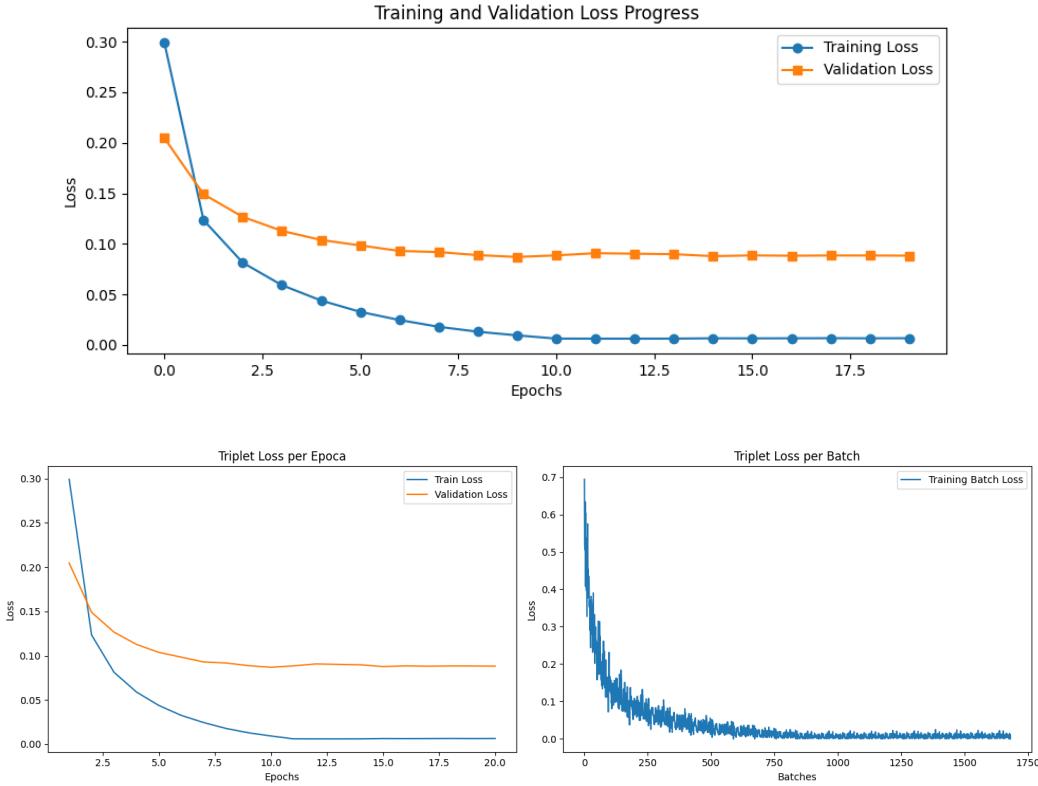
La seconda fase consiste nell'estrazione delle triplette con l'obiettivo di garantire una

maggior divisione tra immagine reale ed immagine fake. Le triplette identificate sono rispettivamente: anchor, positive e negative. In anchor sono presenti le immagini reali, in positive sono presenti le immagini che presentano una similarità con le immagini reali (target=0) ed infine in negative sono presenti le immagini fake (target=1).

La terza fase consiste nel passare gli embedding generati nella fase precedente e le rispettive triplette, organizzate in batch di dimensione fissa (ovvero 128), alla rete neurale Triplet-Model che utilizza la Triplet Loss per regolarizzare i pesi delle triplette sulla rete, massimizzando la distanza tra anchor e negative e minimizzando la distanza tra anchor e positive.

Sulla base delle varie fasi appena descritte, sono state effettuate queste analisi sia per il dataset Cycle-GAN, dataset Pro-GAN che per entrambi i dataset insieme con lo scopo di visualizzarne i comportamenti. In seguito, riportiamo le analisi effettuate per il dataset Cycle-GAN:

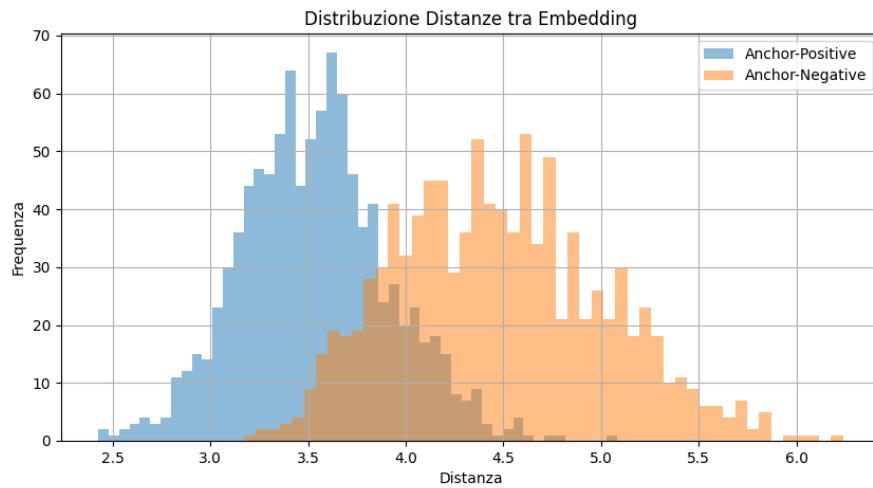
I seguenti grafici mostrano rispettivamente train e validation-loss finale (indica quanto il modello performa bene sui dati) in seguito ad un addestramento in 20 epoche. L'addestramento termina a partire dalla quinta epoca, dal momento che la validation-loss ha raggiunto il miglior risultato possibile.



Come si può notare dai grafici, la training loss scende rapidamente e si stabilizza a un valore basso (circa 0.05), mentre la validation loss inizialmente scende (fino a circa la seconda epoca) ma, successivamente, risale leggermente e si stabilizza a un valore più alto (circa

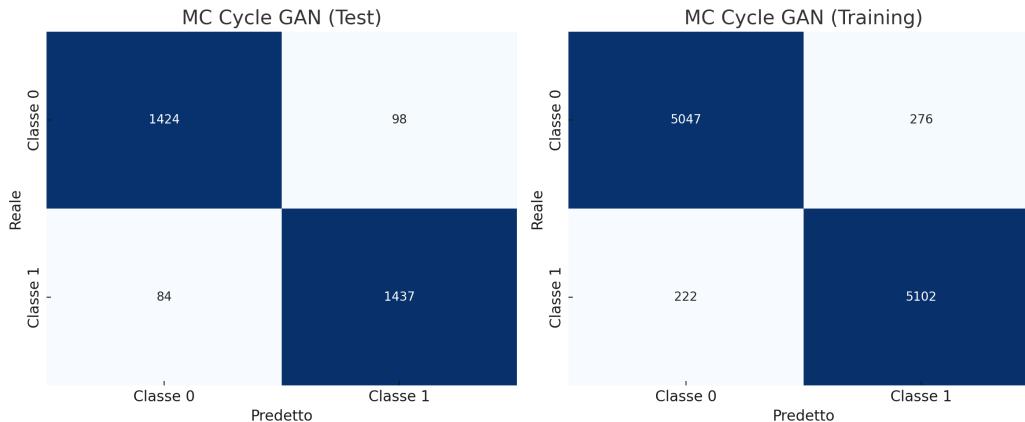
0.15) rispetto alla training loss e il gap tra training e validation loss continua ad aumentare fino a stabilizzarsi con una differenza di circa 0.10 (0.15-0.05).

Il seguente grafico rappresenta le distanze tra le triplette per i vettori di embeddings, rispettivamente le distanze tra anchor-positive e anchor negative. Il grafico suggerisce una netta separazione tra anchor-positive e anchor-negative, confermando in questo modo una buona separazione tra immagini real e fake.



Per l'addestramento del modello CLIP è stato utilizzato SVM (Support Vector Machine) i cui dati di addestramento sono stati suddivisi in 70% per il training, 20% per il Test Set e il restante 10% per Validation Set.

L'analisi è stata effettuata prendendo in considerazione 15210 campioni, dei quali 10647 sono stati selezionati per il Training Set, 1520 per la Validation Set ed infine 3043 per il Test Set. Per ogni set di dati, è stata riportata nel grafico seguente le rispettive matrici di confusione.



Per quanto riguarda il calcolo delle metriche di Accuracy, Precision, Recall, F1-Score, è

stata riportata nella tabella sottostante, la percentuale di accuracy raggiunta da ogni metrica:

Metriche	Percentuale
Accuracy	95.3%
Precision	95.3% -Media tra Precision Classe 0 = 95.79% e Precision Classe 1 = 94.87%
Recall	95.3% -Media tra Recall Classe 0 = 94.81% e Recall Classe 1 = 95.83%
F1-Score	95.3% -Media tra F1-Score Classe 0 = 95.30% e F1-Score Classe 1 = 95.35%

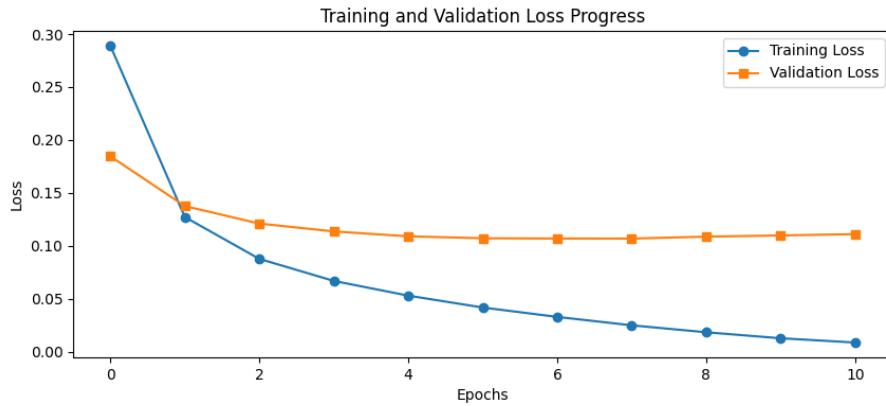
Table 1: Lista delle Metriche con la rispettiva accuracy raggiunta per il Training

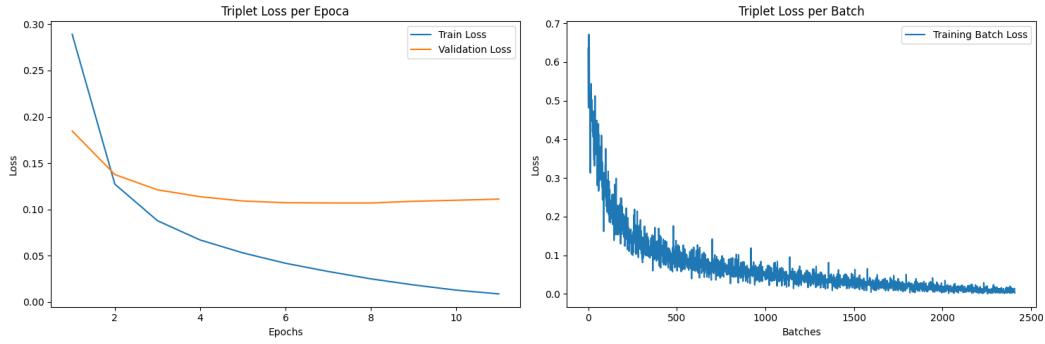
Metriche	Percentuale
Accuracy	94%
Precision	94% -Media tra Precision Classe 0 = 94.43% e Precision Classe 1 = 93.62%
Recall	94% -Media tra Recall Classe 0 = 93.56% e Recall Classe 1 = 94.48%
F1-Score	94% -Media tra F1-Score Classe 0 = 93.99% e F1-Score Classe 1 = 94.04%

Table 2: Lista delle Metriche con la rispettiva accuracy raggiunta per il Test Set

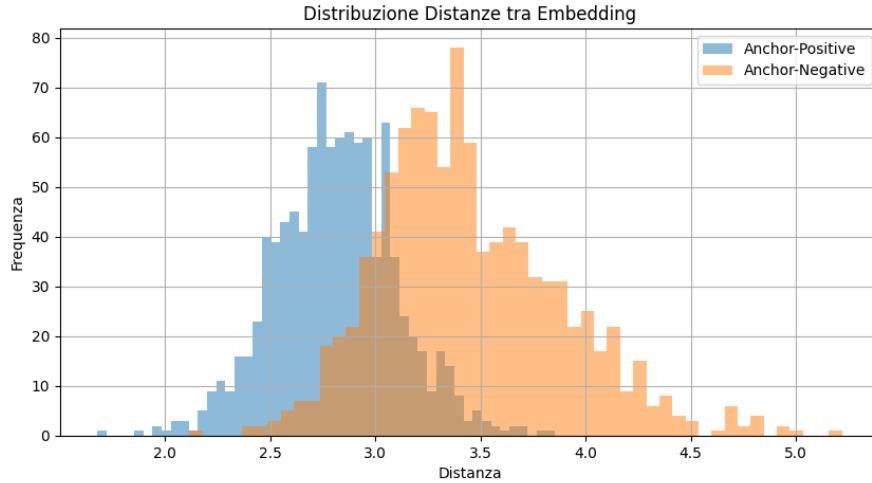
Segue l'analisi effettuata per il dataset Pro-GAN:

Anche per il dataset Pro-GAN, sono stati riportati i grafici di train e validation loss per valutare quanto il modello riesce a performare bene sui dati. Dai grafici seguenti si può notare che sia train che validation-loss seguono lo stesso andamento, stabilizzandosi entrambe nelle epoche finali, con un gap minore di 0.03 e suggerendo un leggero overfitting. In conclusione, il grafico riesce a generalizzare bene sui dati a disposizione.



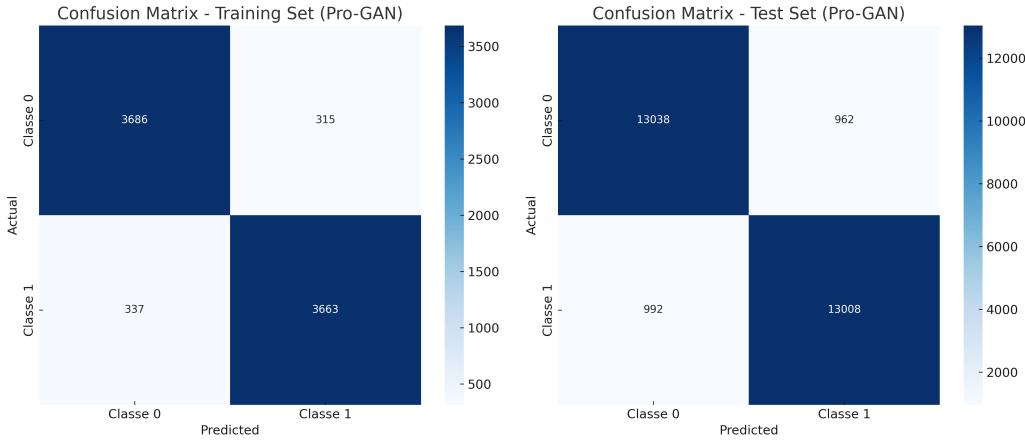


Il seguente grafico rappresenta le distanze tra le triplette per i vettori di embeddings rispettivamente le distanze tra anchor-positive e anchor-negative. Il grafico seguente mostra una netta separazione tra quest'ultimi, confermando una buona separazione tra immagini real e fake.



Anche in questo caso per l'addestramento del modello CLIP è stato utilizzato SVM (Support Vector Machine) i cui dati di addestramento sono stati suddivisi in 70-20-10 come per il dataset Cycle-GAN.

L'analisi è stata effettuata prendendo in considerazione 40000 campioni, dei quali 28000 sono stati selezionati per il Training Set, 3999 per la Validation Set ed infine 8001 per il Test Set. Per ogni set di dati, è stata riportata nel grafico seguente le rispettive matrici di confusione.



Per quanto riguarda il calcolo delle metriche di Accuracy, Precision, Recall, F1-Score, è stata riportata nella tabella sottostante, la percentuale di accuracy raggiunta da ogni metrica:

Metriche	Percentuale
Accuracy	93.02%
Precision	93.02% -Media tra Precision Classe 0 = 92.93% e Precision Classe 1 = 93.11%
Recall	93.02% -Media tra Recall Classe 0 = 93.13% e Recall Classe 1 = 92.91%
F1-Score	93.02% -Media tra F1-Score Classe 0 = 93.03% e F1-Score Classe 1 = 93.01%

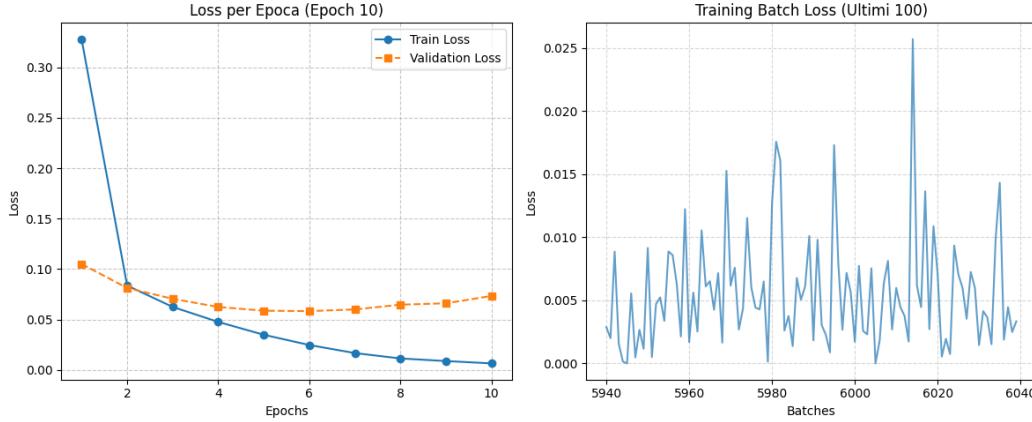
Table 3: Lista delle Metriche con la rispettiva accuracy raggiunta per il Training

Metriche	Percentuale
Accuracy	91.85%
Precision	91.85% -Media tra Precision Classe 0 = 91.62% e Precision Classe 1 = 92.08%
Recall	91.85% -Media tra Recall Classe 0 = 92.13% e Recall Classe 1 = 91.57%
F1-Score	91.85% -Media tra F1-Score Classe 0 = 91.87% e F1-Score Classe 1 = 91.83%

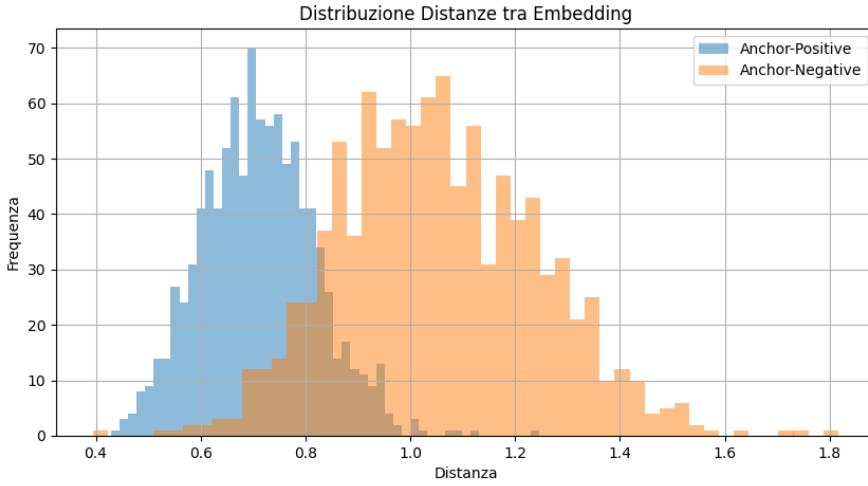
Table 4: Lista delle Metriche con la rispettiva accuracy raggiunta per il Test Set

Segue l'analisi effettuata per il dataset combinato:

Il seguente grafico mostra train e validation-loss finale in seguito ad un addestramento in 20 epoche. Osservando il grafico, è possibile notare come la training-loss diminuisce rapidamente nelle prime epoche e in seguito alla sesta epoca, inizia a stabilizzarsi verso un valore basso (circa 0.05) mentre la validation-loss segue inizialmente l'andamento della training-loss per le prime epoche, ma successivamente si stabilizza dopo la quarta epoca senza migliorare (circa 0.10). In conclusione, è presente overfitting a partire dalla 3-4 epoca.



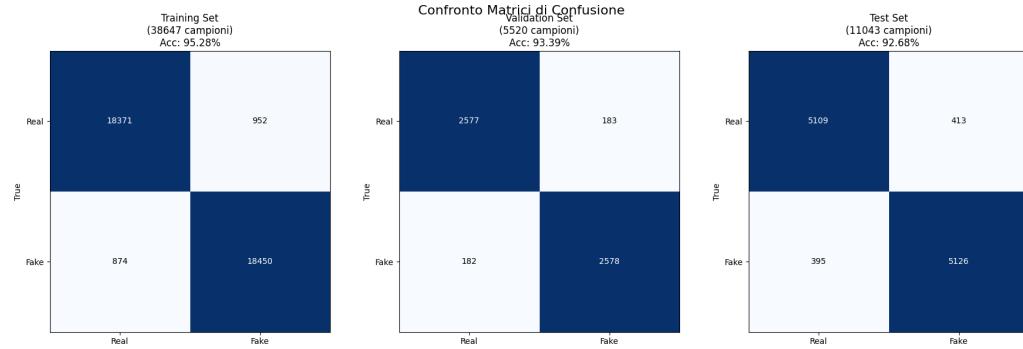
Il seguente grafico rappresenta le distanze tra le triplette per i vettori di embeddings rispettivamente le distanze tra anchor-positive e anchor-negative. Il grafico suggerisce una netta separazione tra quest'ultimi, di conseguenza si verifica una buona separazione tra immagini real e fake.



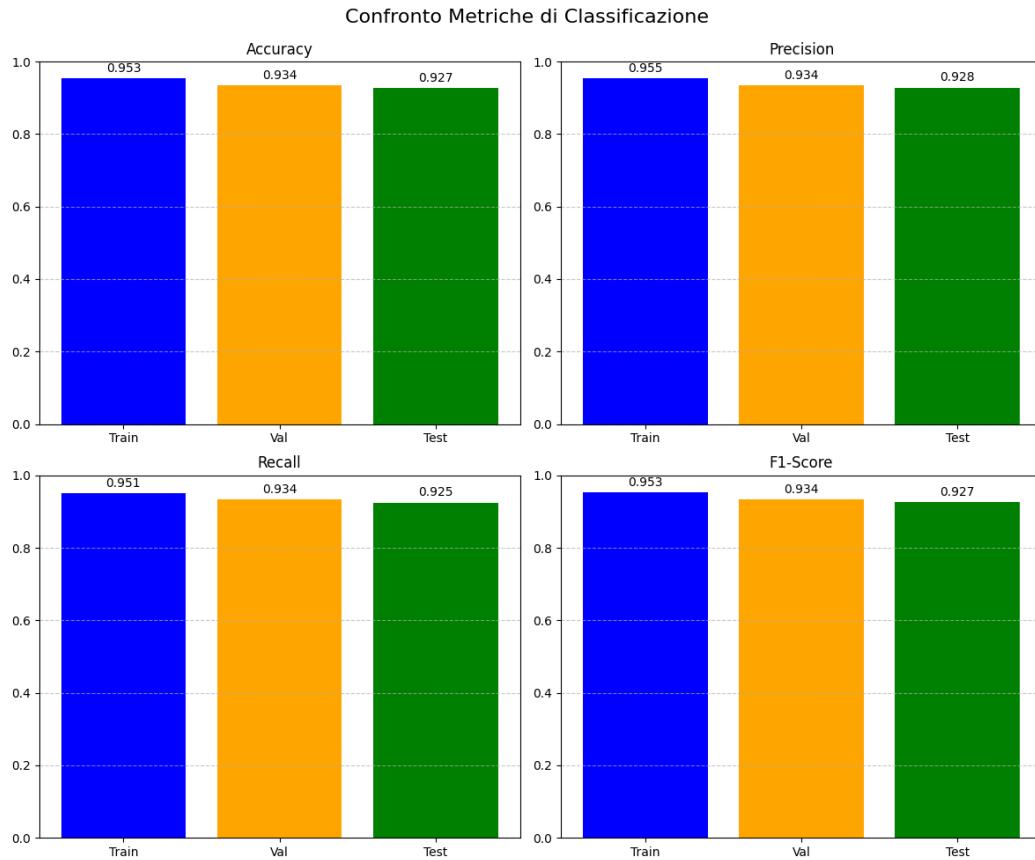
Per l'addestramento del modello CLIP è stato utilizzato SVM (Support Vector Machine) i cui dati di addestramento sono stati suddivisi in 70% per il training, 20% per il Test Set e il restante 10% per Validation Set.

L'analisi è stata effettuata prendendo in considerazione 55210 campioni, dei quali 38647 sono stati selezionati per il Training Set, 5520 per la Validation Set ed infine 11043 per il

Test Set. Per ogni set di dati, è stata riportata nel grafico seguente le rispettive matrici di confusione con la relativa accuracy raggiunta per ognuna di esse.

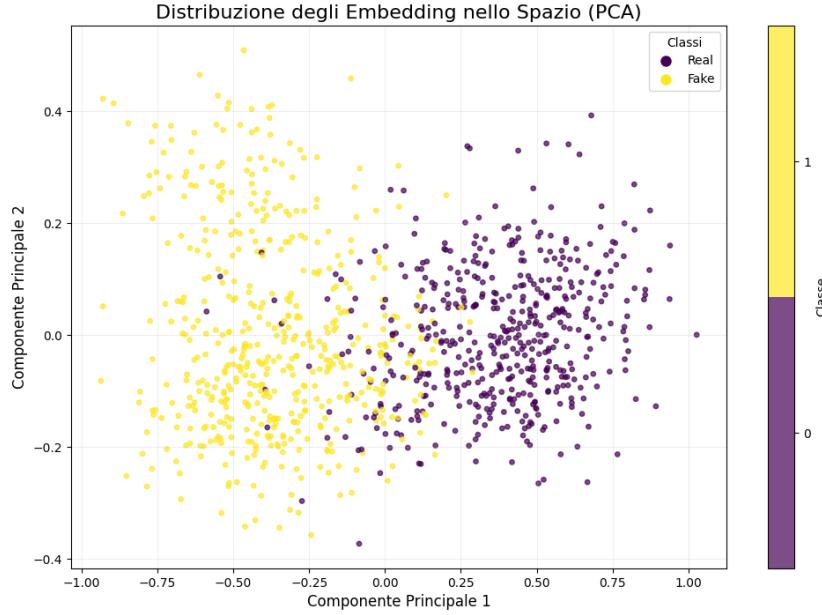


Il seguente grafico mostra le metriche Accuracy, Precision, Recall e F1-Score calcolate per Training, Validation e Test Set con la relativa accuratezza raggiunta.



Il grafico successivo, mostra la distribuzione degli embeddings per le immagini reali e fake utilizzando la PCA come algoritmo di clustering con l'obiettivo di visualizzare, in seguito a tutte le analisi effettuate precedentemente, quanto bene il modello riesce a differenziare le immagini reali dalle immagini fake. La PCA è stata effettuata sui dati di testing per una rappresentazione più chiara dei dati e suggerisce una buona separazione tra i due cluster,

anche se sono presenti, in alcuni punti, elementi che si mischiano tra loro.



3.2 Riconoscimento di Immagini Real e Fake tramite l'estrazione delle informazioni di Immagini + Testo

In questa sezione invece, ci si concentra principalmente su come sono state acquisite le immagini ed il testo (feature category presente nel dataset) utilizzando il modello CLIP (Contrastive Language-Image Pretraining) che, a sua volta, carica un transformer Vit-B/16 e come quest'ultimo è stato addestrato per farsi che tale modello sia in grado di fare distinzione tra immagini reali e fake.

Per quanto riguarda la progettazione, le operazioni che sono state effettuate sono le stesse, con la differenza che, in questo caso, oltre ad estrarre il vettore di embedding per le immagini di dimensione 512, è stato estratto il vettore di embedding per il testo (category) anch'esso di dimensione pari a 512. Inoltre per quest'ultimo, viene creato un dizionario che consente di associare ad ogni embedding dell'immagine estratto il corrispondente embedding del testo. Entrambi i vettori successivamente vengono ridimensionati a 128 con lo scopo di proiettare sia l'embedding dell'immagine che l'embedding del testo in uno spazio di dimensione ridotta.

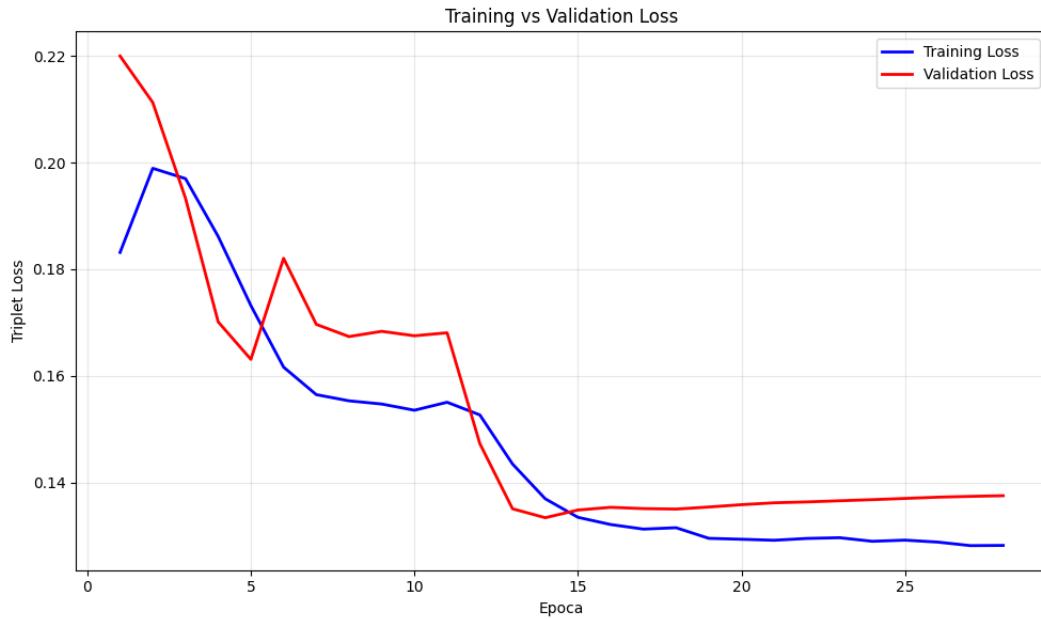
A tal proposito, sono stati utilizzati due approcci:

1. **Somma tra embeddings:** somma elemento per elemento tra le due proiezioni.
2. **Prodotto degli embedding:** prodotto elemento per elemento tra le due proiezioni con lo scopo di cogliere l'interazione diretta tra immagine e testo.

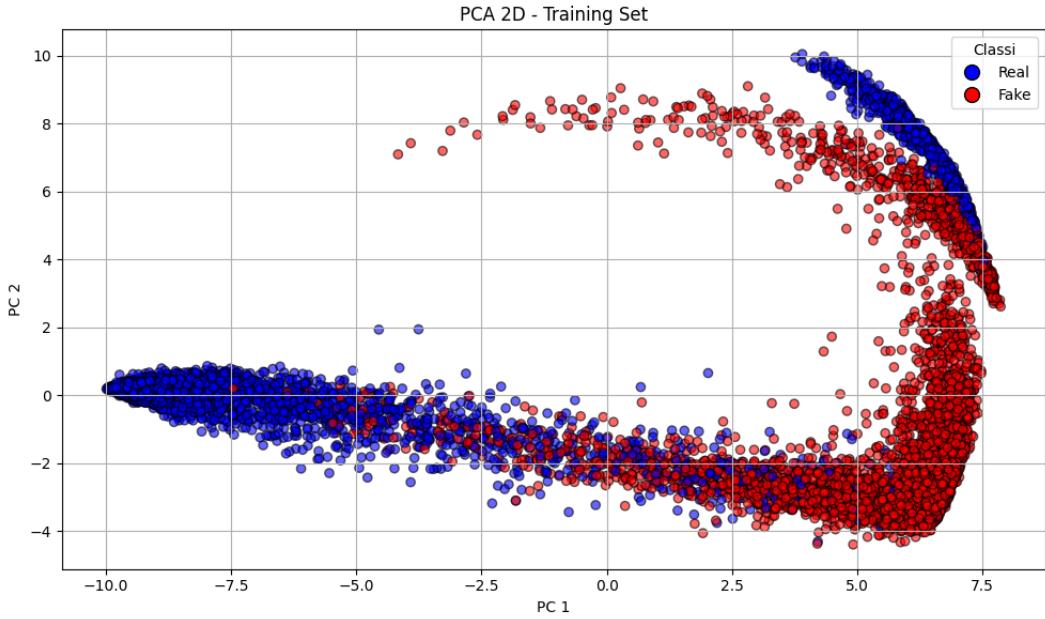
Dal momento che il modello CLIP non può ricevere più vettori contemporaneamente, questi due vettori contenente le due informazioni estratte, sono stati fusi insieme formando un unico vettore, definito come vettore multimodale di dimensione pari a 1024, normalizzato successivamente a 128.

Sulla base della premessa fatta precedentemente, in questa sezione verranno visualizzati i risultati ottenuti effettuando la somma tra embeddings. Le seguenti analisi sono state effettuate sia per il dataset Cycle-GAN, sia per il dataset Pro-GAN che per entrambi i dataset insieme, con lo scopo di visualizzarne i comportamenti. In seguito, riportiamo le analisi effettuate per il dataset Cycle-GAN:

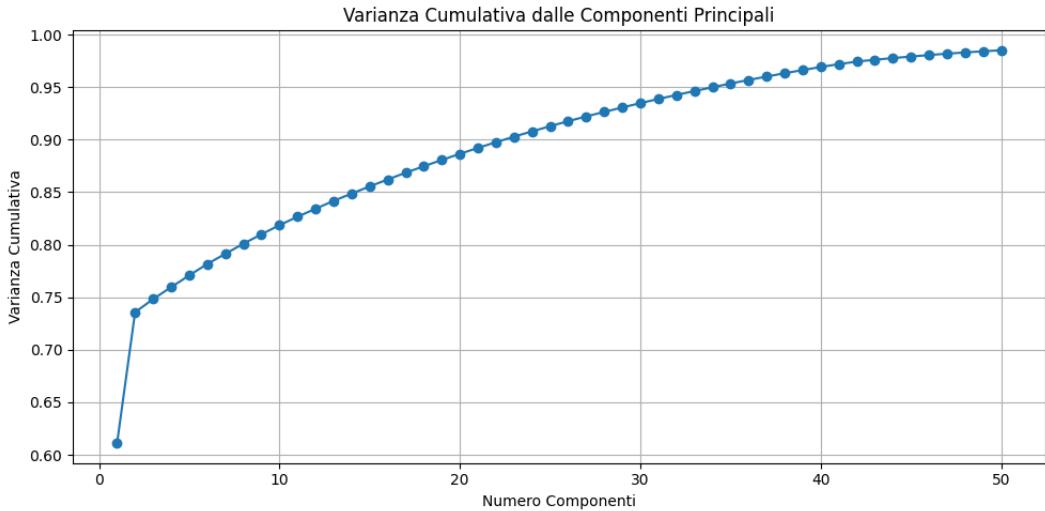
Il seguente grafico mostra l'andamento di training e validation-loss durante l'addestramento di un modello Cycle-GAN (usato per traduzioni tra immagini e testo). Da questo grafico si può notare che sia la training che la validation-loss presentano lo stesso andamento per tutte le epoche, in quanto entrambe si stabilizzano a 0.14, indicando che non vi è presenza di overfitting. Di conseguenza il modello generalizza bene sui dati.



Il seguente grafico, invece mostra la distribuzione degli embeddings per le immagini reali e fake utilizzando la PCA come algoritmo di clustering con l'obiettivo di visualizzare quanto bene il modello riesce a differenziare le immagini reali dalle immagini fake. La PCA è stata effettuata sui dati di training e suggerisce una buona separazione tra immagini reali ed immagini fake, anche se sono presenti alcuni elementi che continuano a mischiarci tra di loro.



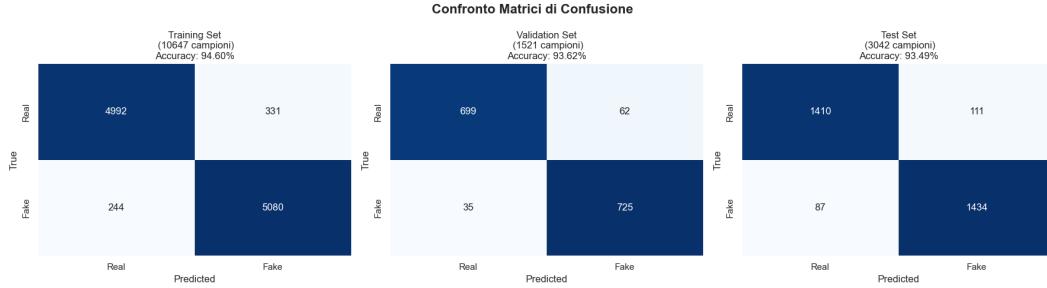
In seguito al grafico della PCA, è stato effettuato il grafico della varianza che consente di verificare quanta variabilità dei dati viene catturata da ogni componente principale. In questo caso, è possibile notare che già con la scelta di 10 componenti, viene conservata l'80% dell'informazione, suggerendo un'alta varianza dei dati determinata dal fatto che, precedentemente, era stata confermata l'assenza di overfitting.



Per l'addestramento del modello CLIP è stato utilizzato SVM (Support Vector Machine) i cui dati di addestramento sono stati suddivisi in 70% per il training, 20% per il Test Set e il restante 10% per Validation Set.

L'analisi è stata effettuata prendendo in considerazione 15210 campioni, dei quali 10647 sono stati selezionati per il Training Set, 1521 per la Validation Set ed infine 3042 per il

Test Set. Per ogni set di dati, è stata riportata nel grafico seguente le rispettive matrici di confusione con la relativa accuracy raggiunta per ognuna di esse.



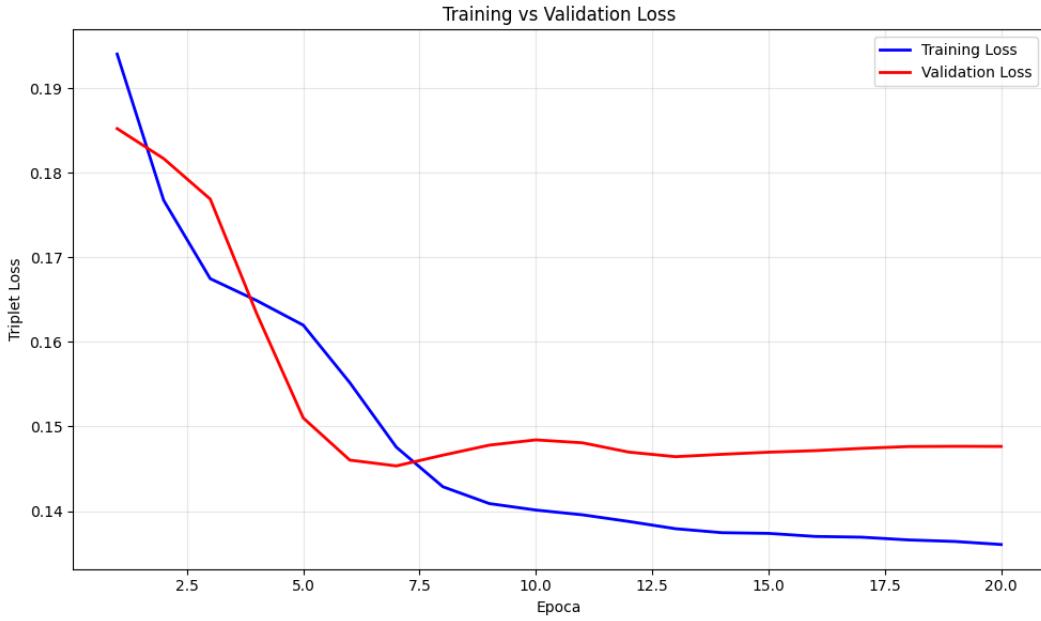
Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score calcolate per il dataset Cycle-GAN con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	94.61%	93.68%	93.50%
Recall	94.60%	93.62%	93.49%
F1-Score	94.60%	93.62%	93.49%

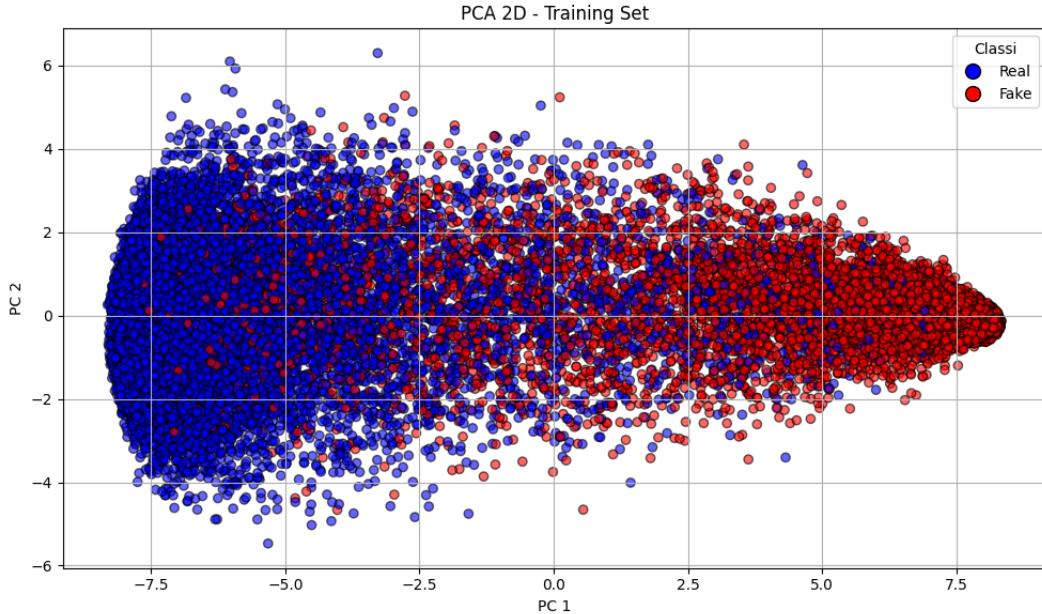
Table 5: Lista delle Metriche

Segue l'analisi effettuata per il dataset Pro-GAN:

Il seguente grafico descrive l'andamento di training e validation-loss per il dataset Pro-gan. Anche in questo caso, come per il dataset Cycle-GAN, vi è assenza di overfitting dato che, sia train che validation decrescono insieme contemporaneamente, difatti all'epoca 20, entrambe convergono allo stesso valore, ovvero 0.14. Nelle ultime epoche (15-20), entrambe le curve si stabilizzano gradualmente, indicando che il modello ha raggiunto la sua massima capacità di apprendimento sui dati.

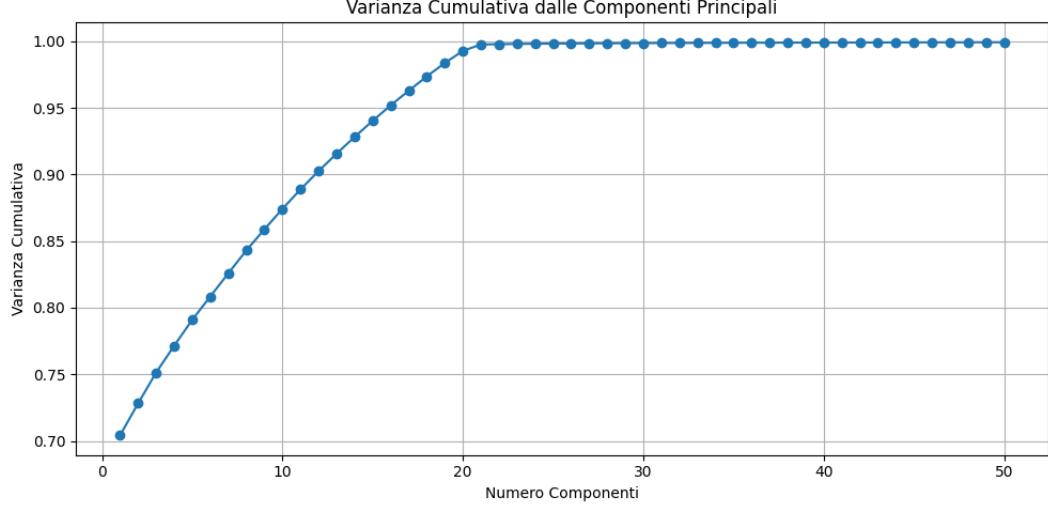


Anche in questo caso, come per il dataset Cycle-GAN, è stato effettuato il grafico della PCA che mostra la distribuzione degli embeddings per le immagini reali e fake. Anche in questo caso la PCA è stata effettuata sui dati di training e suggerisce una buona separazione tra immagini reali ed immagini fake, nonostante siano presenti alcuni elementi che continuano a mischiararsi tra di loro.

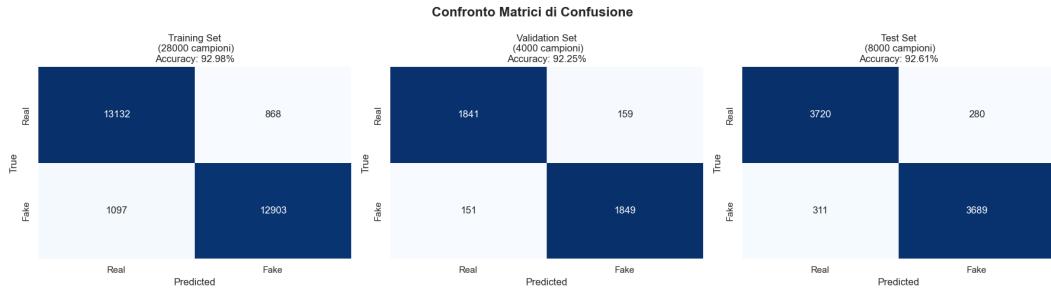


Sulla base del grafico della PCA analizzato precedentemente, è stato effettuato il grafico che rappresenta la varianza tra le componenti principali. Come suggerisce il grafico sottostante, la prima componente è in grado di spiegare circa il 70.5% della varianza totale del dataset, che di per sé, rappresenta un valore molto alto, suggerendo una forte direzione di

variazione nei dati. Aggiungendo le componenti successive, la varianza cumulativa cresce molto rapidamente, infatti con sole 10 componenti si riesce a spiegare quasi il 90% della varianza. Dopo le prime 20-22 componenti, si raggiunge il 99% della varianza spiegata e la curva si appiattisce drasticamente raggiungendo il plateau.



Anche per il dataset Pro-GAN come per Cycle-GAN, sono state riportate in seguito le matrici di confusione con le rispettive accuracy ottenute durante la fase di addestramento del modello utilizzando SVM (Support Vector Machine). Inoltre, anche in questo caso, è stata effettuata la stessa suddivisione del set di dati effettuata precedentemente per il dataset Cycle-GAN, con la differenza che, per il Training, sono stati selezionati 28000 campioni, per la Validation 4000 campioni ed infine per il Test Set 8000 campioni.



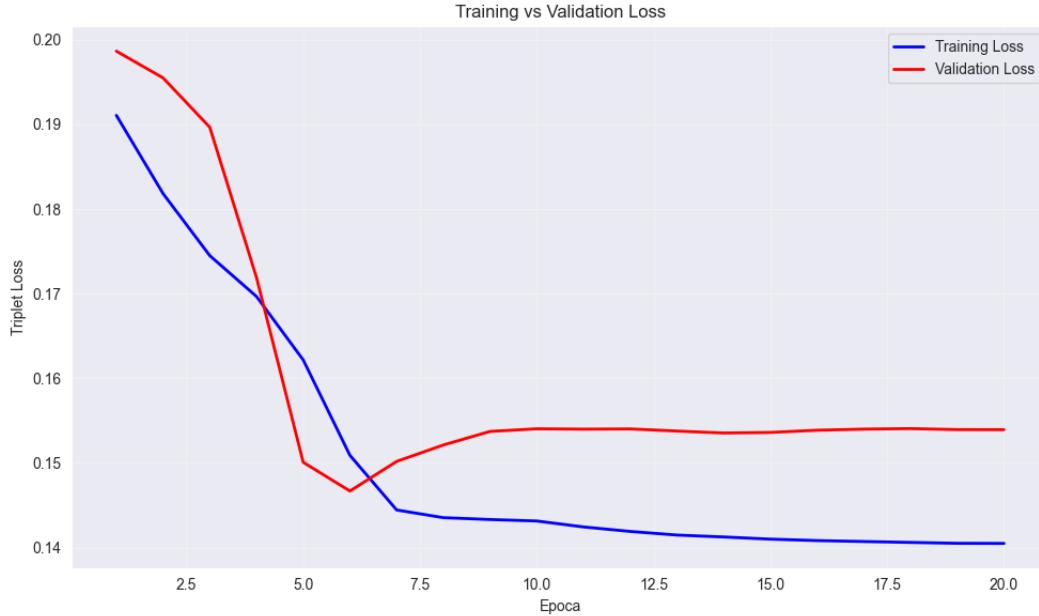
Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score calcolate per il dataset Pro-GAN con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	92.99%	92.25%	92.62%
Recall	92.98%	92.25%	92.61%
F1-Score	92.98%	92.25%	92.61%

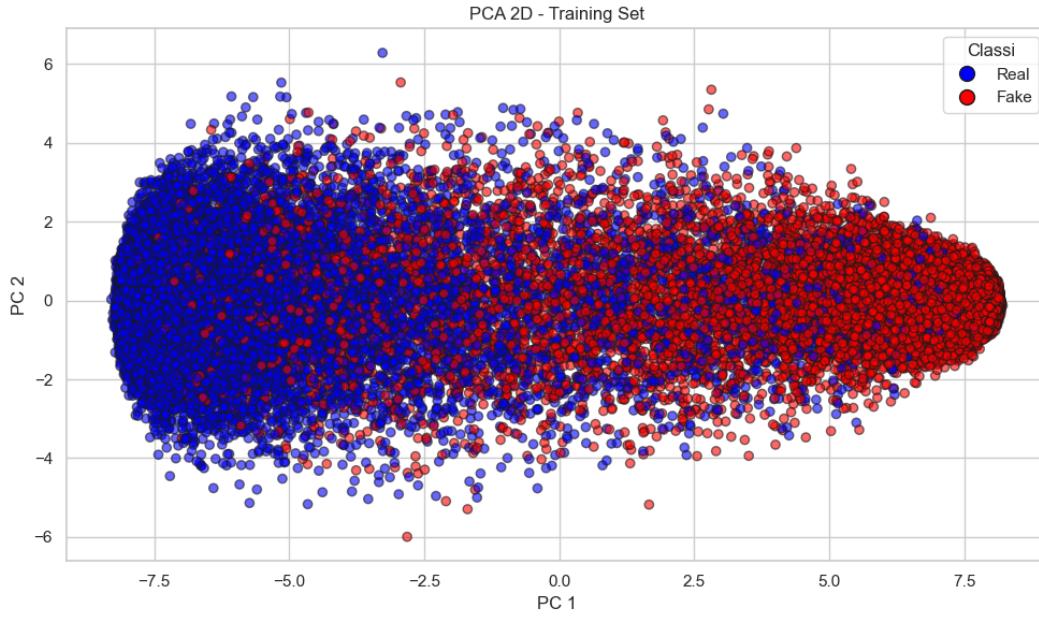
Table 6: Lista delle Metriche

Segue l'analisi effettuata per il dataset combinato:

L'immagine seguente mostra l'andamento di train e validation-loss. Come si può evin-
cere dal grafico, la training-loss scende sempre più velocemente, indicando che il modello
sta apprendendo efficacemente dai dati di addestramento. Anche la validation-loss segue
un andamento simile alla training-loss fino alla sesta epoca, ma successivamente inizia ad
aumentare, indicando che il modello ha iniziato a sovradattarsi ai dati di training. Questo
suggerisce una buona geeralizzazione sui dati.

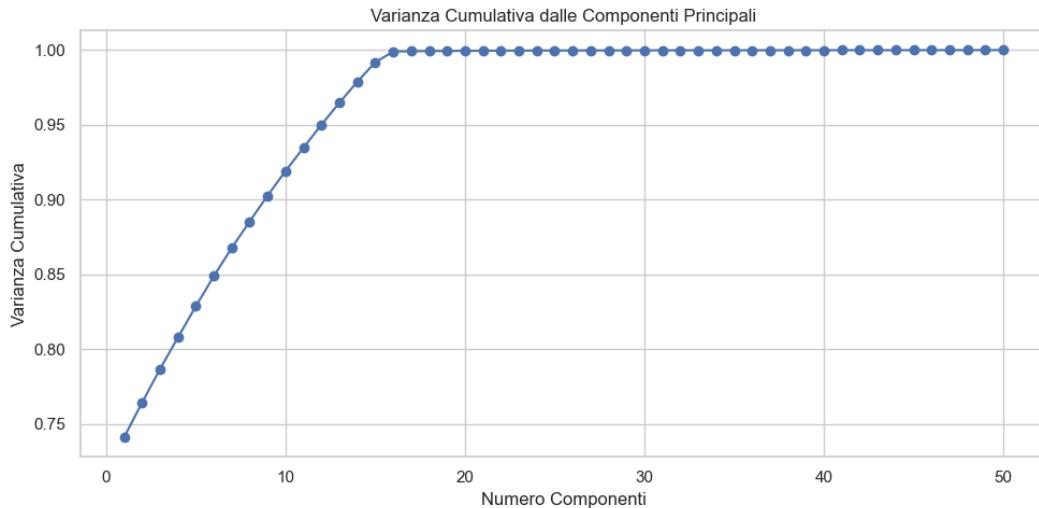


Come per i due dataset Cycle-GAN e Pro-GAN, anche per il dataset combinato è stato
effettuato il grafico della PCA per verificare quanto bene riesce a fare distinzione tra im-
magine reale ed immagine fake dopo l'addestramento del modello. Il grafico sottostante
suggerisce una buona separazione tra immagine reale ed immagine fake, anche se sono pre-
senti alcune componenti che si mischiano ta di loro. Anche in questo caso, come negli altri
due casi precedenti, la PCA è stata effettuata sui dati di Training.



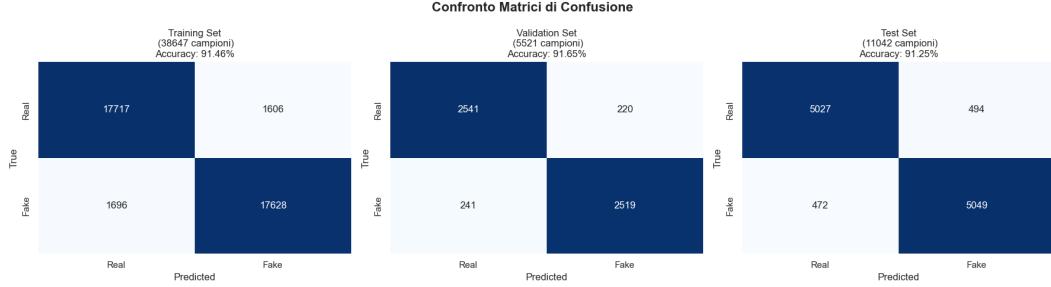
In seguito al grafico della PCA, nel grafico sottostante è stato rappresentato il grafico della varianza, il cui scopo è comprendere quante componenti principali sono necessarie per rappresentare l'informazione contenuta all'interno del dataset, senza dover utilizzare tutte le variabili originali. Dal seguente grafico è possibile notare che, con una sola componente principale, si riesce già a spiegare circa il 74% della varianza totale del dataset. Aggiungendo le prime 10 componenti, la varianza cumulativa spiegata supera il 90%.

Dopo circa 17 componenti, la curva diventa quasi piatta e raggiunge il valore di 1.0 (100%). Questo significa che le prime 17 componenti sono sufficienti per catturare la quasi totalità dell'informazione presente nel dataset originale, mentre le componenti successive risultano ridondanti.



Anche per il dataset combinato, sono state riportate in seguito le matrici di confusione

con le rispettive accuracy ottenute durante la fase di addestramento del modello. Inoltre, anche in questo caso, è stata effettuata la stessa suddivisione del set di dati effettuata precedentemente (70-20-10) per gli altri due dataset presi singolarmente, con la differenza che, per il Training, sono stati selezionati 38647 campioni, per la Validation 5521 campioni ed infine per il Test Set 11042 campioni.



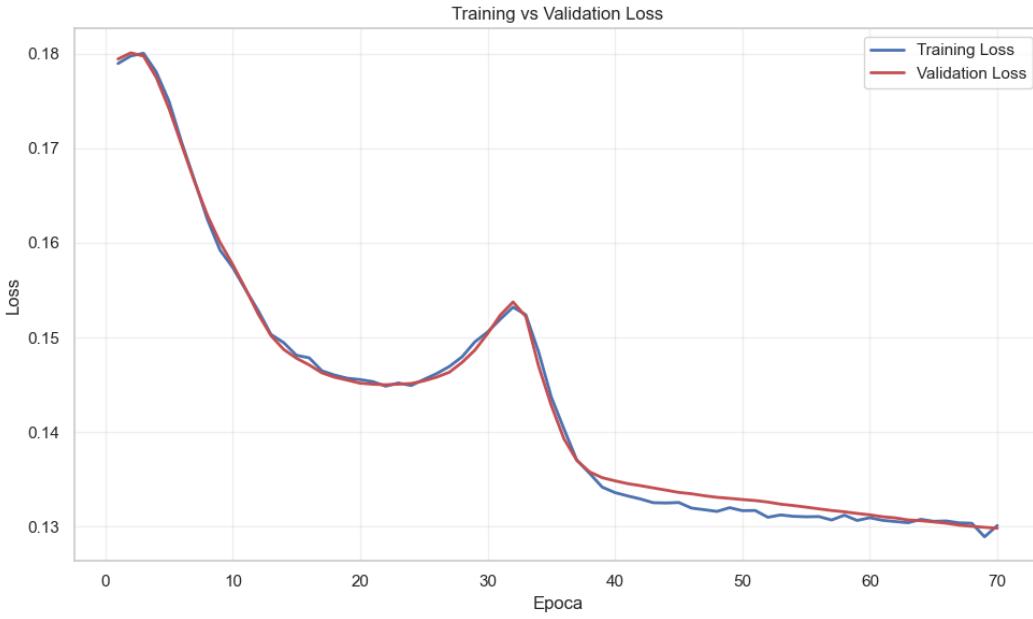
Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	91.46%	91.65%	91.25%
Recall	91.46%	91.65%	91.25%
F1-Score	91.46%	91.65%	91.25%

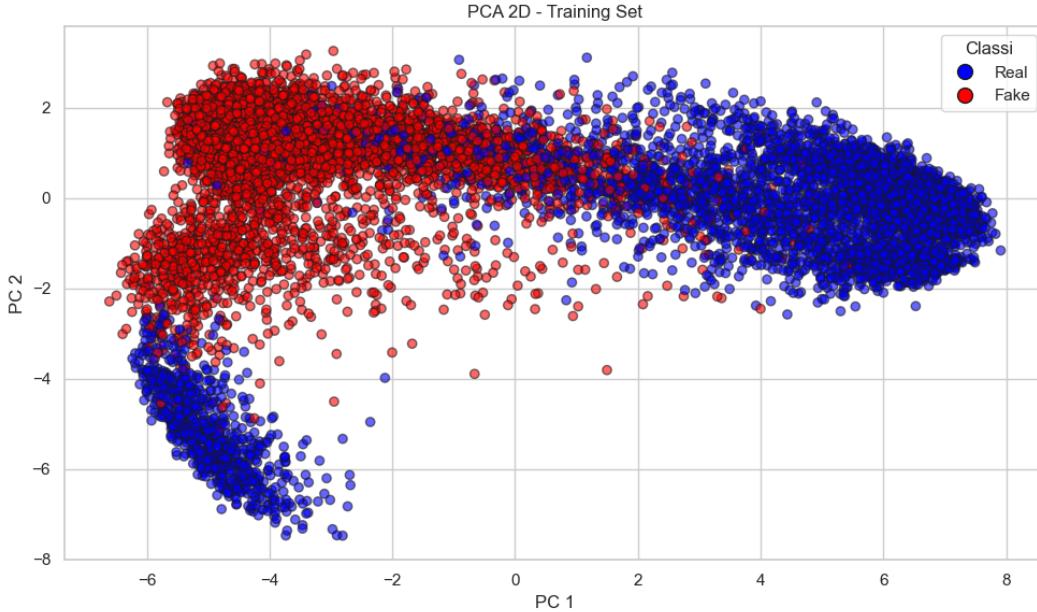
Table 7: Lista delle Metriche

In questa sezione verranno visualizzati i risultati ottenuti effettuando il prodotto tra le singole componenti contenute negli embeddings. Nello specifico, vengono estratti singolarmente l'embedding dell'immagine, l'embedding del testo e la combinazione tra i due (128x3=384) e successivamente, il vettore finale ottenuto viene normalizzato ad una dimensione pari a 128. Le seguenti analisi sono state effettuate sia per il dataset Cycle-GAN, sia per il dataset Pro-GAN che per entrambi i dataset insieme, con lo scopo di visualizzarne i comportamenti. In seguito, riportiamo le analisi effettuate per il dataset Cycle-GAN:

Il seguente grafico mostra l'andamento di train e validation-loss. E' possibile notare che train e validation-loss presentano entrambe lo stesso andamento. Inizialmente (0-20 epoche) vi è una rapida diminuzione di entrambe le loss da 0.18 a 0.145, ma successivamente, si osserva un improvviso aumento (epoche 30-32) delle loss fino a 0.153, seguito da un rapido recupero. Le loss si stabilizzano intorno a 0.13, con training e validation loss che rimangono molto vicine

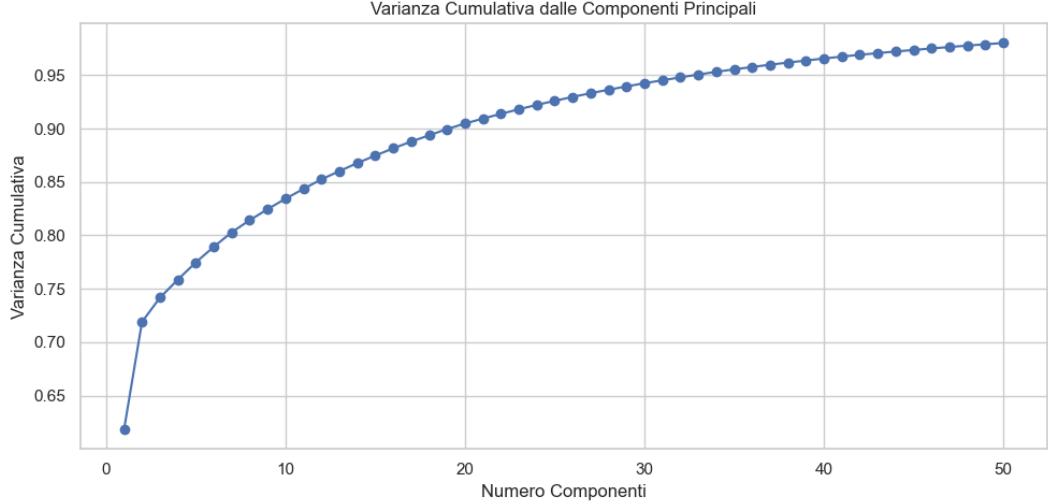


Il seguente grafico, invece mostra la distribuzione degli embeddings per le immagini reali e fake utilizzando la PCA come algoritmo per la distinzione tra real e fake. La PCA è stata effettuata sui dati di training e suggerisce una buona separazione tra immagini reali ed immagini fake, anche se sono presenti alcuni elementi che continuano a mischiarci tra di loro.

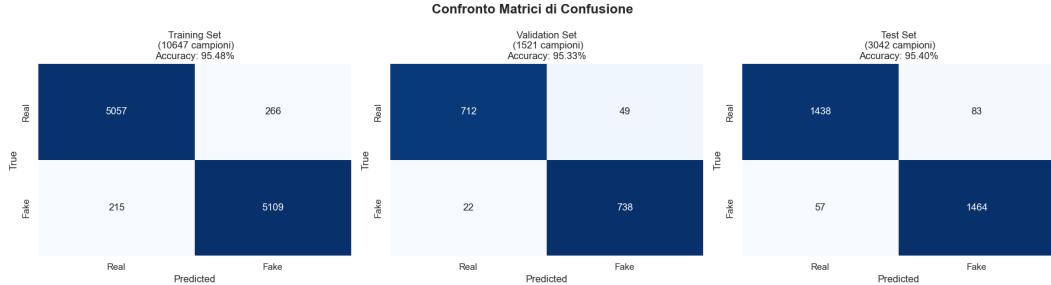


Come conseguenza al grafico della PCA, è stato effettuato il grafico della varianza cumulativa delle componenti principali. Come suggerisce il grafico sottostante, la prima componente spiega circa il 65% della varianza, con un incremento graduale e uniforme lungo tutte le componenti fino a raggiungere quasi il 100% della varianza cumulativa verso la componente 50. Non ci sono singole componenti dominanti, di conseguenza l'informazione

è distribuita abbastanza uniformemente.



Il grafico sottostante rappresenta le matrici di confusione per Training, Validation e Test con la relativa accuratezza ottenuta in seguito all’addestramento effettuato. L’analisi è stata effettuata prendendo in considerazione 15210 campioni, dei quali 10647 sono stati selezionati per il Training Set, 1521 per la Validation Set ed infine 3042 per il Test Set.



Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

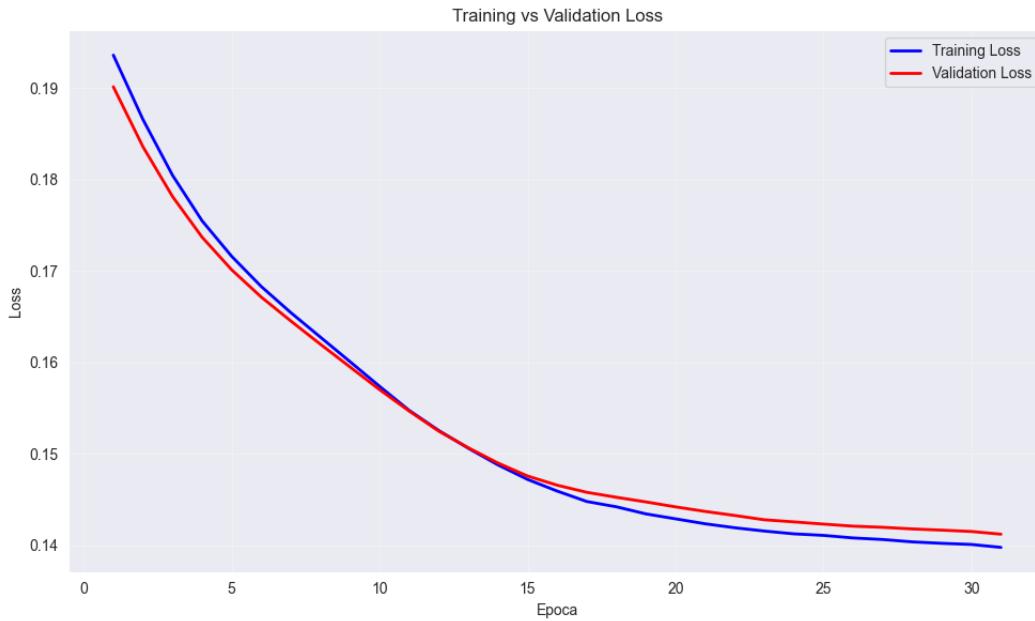
Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	95.49%	95.39%	95.41%
Recall	95.48%	95.33%	95.40%
F1-Score	95.48%	95.33%	95.40%

Table 8: Lista delle Metriche

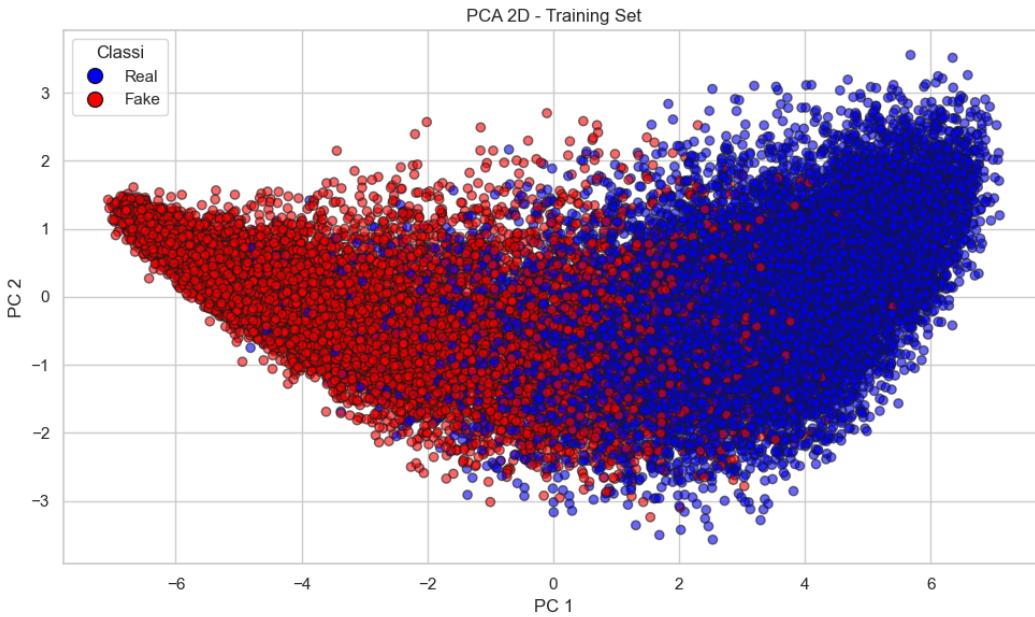
Segue la stessa analisi effettuata per il dataset Pro-GAN:

Il grafico seguente analizza l’andamento di train e validation-loss. Dal grafico si evince che non vi è presenza di overfitting, poichè entrambe le loss diminuiscono costantemente da 0.195 a 0.14, difatti le curve presentano un andamento parallelo e verso le ultime epoche,

le loss si appiattiscono, indicando convergenza.

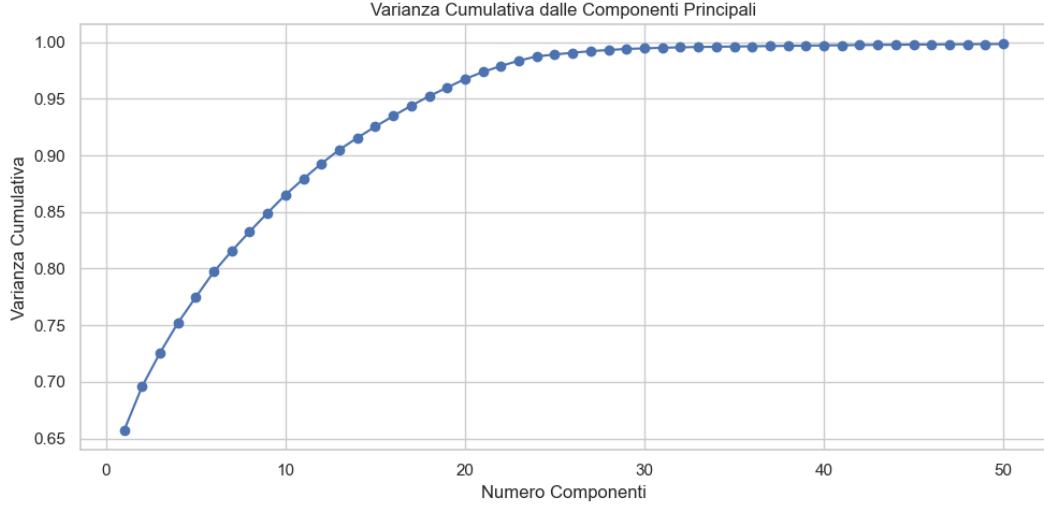


Il seguente grafico, invece mostra la distribuzione degli embeddings per le immagini reali e fake utilizzando la PCA come algoritmo. La PCA è stata effettuata sui dati di training e suggerisce una buona separazione tra immagini reali ed immagini fake, anche se sono presenti alcuni elementi che continuano a mischiarci tra di loro.

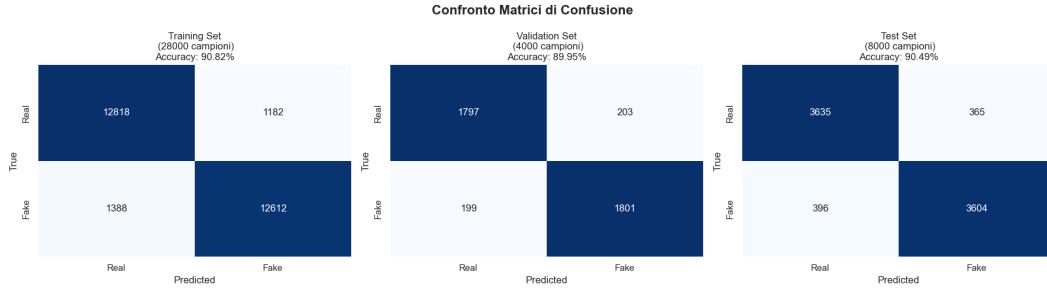


Anche in questo caso, come conseguenza al grafico della PCA, è stato effettuato il grafico della varianza cumulativa delle componenti principali. Come suggerisce il grafico sottostante, la prima componente spiega solo circa il 62% della varianza, seguito da un incremento più

pronunciato tra la prima e la seconda componente. Il comportamento e l'andamento assunto è molto simile al comportamento assunto per il dataset Cycle-GAN ma leggermente più graduale. Anche in questo caso, il seguente grafico raggiunge quasi il 100% della varianza cumulativa.



Il grafico sottostante rappresenta le matrici di confusione per Training, Validation e Test con la relativa accuratezza ottenuta in seguito all'addestramento effettuato utilizzando SVM. L'analisi è stata effettuata prendendo in considerazione 40000 campioni, dei quali 28000 sono stati selezionati per il Training Set, 4000 per la Validation Set ed infine 8000 per il Test Set.



Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	90.83%	89.95%	90.49%
Recall	90.82%	89.95%	90.49%
F1-Score	90.82%	89.95%	90.49%

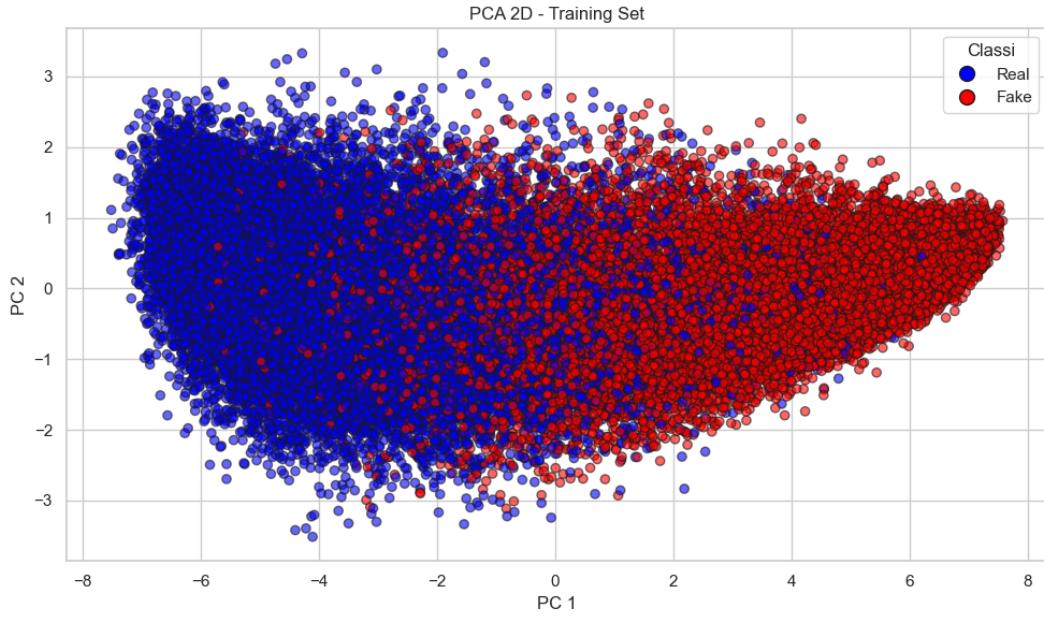
Table 9: Lista delle Metriche

Seguono, infine, le analisi effettuate per i due dataset insieme:

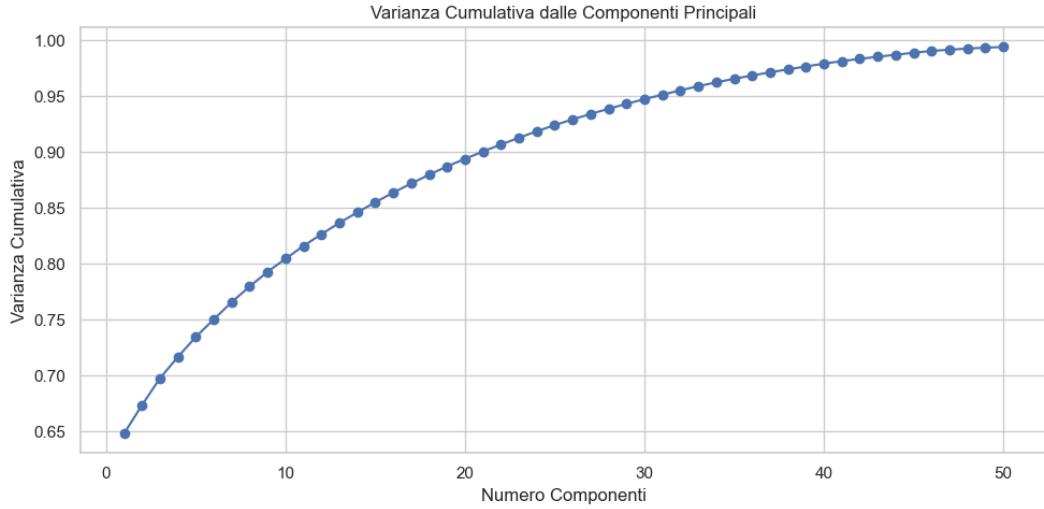
Il seguente grafico mostra il comportamento e l'andamento di train e validation-loss. Dal grafico si evince che, il comportamento assunto è simile al grafico di train e validation-loss di ProGAN ma con alcune differenze. E' possibile notare che, per le epoche iniziali vi sono delle leggere oscillazioni seguito da una rapida discesa per entrambe le loss, come nel grafico per ProGAN fino ad arrivare alle epoche finali, in cui le loss finali risultano essere comparabili, assumendo lo stesso valore (0.142)



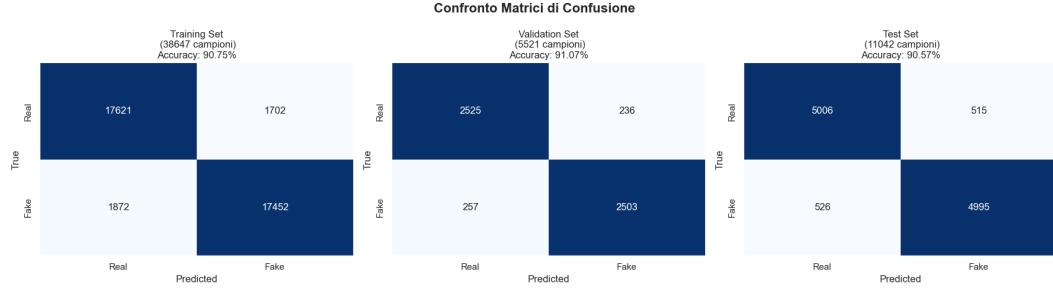
Come per i due dataset Cycle-GAN e Pro-GAN, anche per il dataset combinato è stato effettuato il grafico della PCA per verificare quanto bene riesce a fare distinzione tra immagine reale ed immagine fake dopo l'addestramento del modello. Il grafico sottostante, suggerisce una buona separazione tra immagine reale ed immagine fake, anche se sono presenti alcune componenti che si mischiano tra di loro. Anche in questo caso, come negli altri due casi precedenti, la PCA è stata effettuata sui dati di Training.



Anche in questo caso, è stato effettuato il grafico che mostra la varianza cumulativa delle componenti principali. Come si può evincere dal grafico sottostante, la prima componente raggiunge il 65% della varianza cumulativa, come avveniva per il dataset Cycle-GAN ma, successivamente, raggiunge il 100% della varianza già intorno alla componente 25-30 mentre le componenti successive non aggiungono varianza, rimanendo piatte al 100%.



Il grafico sottostante rappresenta le matrici di confusione per Training, Validation e Test con la relativa accuratezza ottenute in seguito all'addestramento effettuato utilizzando SVM. L'analisi è stata effettuata prendendo in considerazione 55210 campioni, dei quali 38647 sono stati selezionati per il Training Set, 5521 per la Validation Set ed infine 11042 per il Test Set.



Nella tabella sottostante, sono state riportate le metriche Precision, Recall ed F1-Score con la rispettiva percentuale di accuracy raggiunta per ognuna di esse:

Metriche	Percentuale Training	Percentuale Validation	Percentuale Test Set
Precision	90.76%	91.07%	90.57%
Recall	90.75%	91.07%	90.57%
F1-Score	90.75%	91.07%	90.57%

Table 10: Lista delle Metriche

4 Conclusioni e Sviluppi futuri

Sulla base di quanto discusso nei capitoli precedenti, possiamo concludere che entrambi gli approcci utilizzati (estrazione delle sole immagini ed estrazione delle immagini+testo) per fare distinzione tra immagini reali e immagini fake hanno portato al raggiungimento di buoni risultati, in quanto entrambi gli approcci effettuano una buona generalizzazione dei dati. Tuttavia, sulla base dei risultati ottenuti, è possibile effettuare ulteriore ricerca al fine di migliorare i risultati raggiunti, per farsi che il modello sia in grado di effettuare una separazione netta tra immagini reali ed immagini fake.

Link di Riferimento

[1] Dataset DeepFake

[2] Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection
2024 IEEE International Workshop on Information Forensics and Security (WIFS) — Au-
tori: Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino and Luisa
Verdoliva University Federico II of Naples, Italy

[3] Synthetic Image Verification in the Era of Generative Artificial Intelligence What
Works and What Isn't There yet

[4] CLIP DeepFake Detection

[5] CLIP Model and the Importance of Multimodal Embeddings

[6] Image-Classification-CLIP

[7] Guida per l'implementazione di Early Stopping in Pytorch