

Confronto tra reti neurali convoluzionali e reti *fully connected* dopo analisi delle componenti principali per la diagnosi di carcinoma duttale invasivo

Elisa Lazzaroni, Noemi Aime

12 giugno 2024

Abstract

L'estrazione di informazioni significative dalle immagini è una sfida fondamentale nell'apprendimento automatico. Le reti neurali convoluzionali (CNN) si sono dimostrate particolarmente efficaci in questo compito, ma affrontano sfide significative legate all'alta dimensionalità dei dati. La *Principal Component Analysis* (PCA) è una tecnica comunemente utilizzata per ridurre la dimensionalità. In particolare, nel campo della diagnosi medica basata su immagini, la riduzione della dimensionalità senza perdita di informazioni critiche potrebbe migliorare l'efficienza dei modelli di *machine learning*. Questo progetto esplora l'uso di PCA per l'analisi e la classificazione di immagini di tessuti cellulari al microscopio, confrontando le prestazioni di un modello allenato sui dati compressi da PCA con le prestazioni della CNN. Per la CNN è stato utilizzato il modello ResNet18 con l'aggiunta di *layer fully-connected*, mentre con la PCA si sono compressi i dati in diverse componenti principali e si è poi utilizzata una rete *shallow* per l'addestramento. I risultati mostrano che, mentre la CNN mantiene una precisione complessiva elevata, l'uso di PCA con una riduzione significativa delle componenti principali comporta un miglioramento nei tempi di addestramento senza una perdita drastica delle prestazioni. Tuttavia, differenze nelle metriche di *precision* e *recall* suggeriscono che la CNN sia più affidabile per l'identificazione accurata delle aree di tessuto malato, sebbene la PCA offra vantaggi significativi in termini di efficienza computazionale.

1 Introduzione

L'analisi di immagini è una disciplina fondamentale nell'ambito dell'apprendimento automatico, che si occupa di estrarre significato, *pattern* e informazioni utili dalle immagini. Le reti neurali convoluzionali (CNN) hanno dimostrato eccezionali capacità di apprendimento per questo tipo di compiti, ma possono essere soggette a problemi di alta dimensionalità dei dati. La *Principal Component Analysis* (PCA) è una tecnica ampiamente utilizzata per ridurre la dimensionalità dei dati, ma il suo impatto sulle prestazioni delle CNN è ancora oggetto di studio.

Uno dei problemi principali quando si vuole impiegare un modello di *machine learning* su un *dataset* fatto di immagini è che queste possono avere dimensioni troppo grandi per essere processate in maniera efficiente. Questo problema si fa particolarmente sentire nell'ambito degli studi su immagini in ambito medico dove un'alta risoluzione è indispensabile per effettuare una buona diagnosi. Poder fare affidamento su una tecnica come la PCA per abbassare la dimensionalità del problema mantenendo lo stesso livello di prestazioni sarebbe

quindi un valido aiuto nello sviluppo di reti neurali per la diagnosi preventiva basata su analisi di immagini. In questo lavoro, perciò, esamineremo l'utilizzo di CNN e PCA per analizzare e classificare immagini relative a *scan* al microscopio di tessuti cellulari acquisite per la diagnosi di carcinoma duttale invasivo (IDC), una delle forme più comuni di tumori al seno, valutando l'impatto sull'efficacia del modello della riduzione delle dimensioni.

2 Preparazione dei dati

Per la parte relativa all'applicazione della CNN si è fatto riferimento al lavoro presentato nel 2014 dal professor Anant Madabhushi, Case Western Reserve University, e il suo gruppo (Angel Cruz-Roa, Ajay Basavanhally, Fabio A. González, Hannah Gilmore, Michael D Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, Anant Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks", Proceedings Volume 9041, Medical Imaging 2014: Digital Pathology; 904103 (2014), DOI: 10.1117/12.2043872).

Il *dataset* utilizzato per svolgere questo studio (Paul Mooney, 2018, Breast Histopathology Images, Version 1, <https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images/data>) è molto simile al *dataset* utilizzato nell'articolo citato. Il *dataset* consiste in immagini di tessuto relativo a 279 pazienti diagnosticati con IDC; come viene anche spiegato nell'articolo, le immagini sono state suddivise in *patch* da 50x50 pixel ciascuna, per un totale di 277524 immagini totali. Essendo una suddivisione di uno scan complessivo, parte di queste immagini conterranno IDC (*label* 1) e parte no (*label* 0). A partire da questo insieme di dati si può quindi costruire una classificazione binaria che in definitiva riesca a distinguere le zone che contengono IDC da quelle sane, riproducendo il lavoro dell'istopatologo che, analizzando le scansioni al microscopio, distingue le zone di tessuto malato da quelle di tessuto sano. La Figura 1 mostra più nel dettaglio, rispettivamente, la distribuzione delle *patch* tra i pazienti, la percentuale di *patch* sane e malate per paziente e il numero complessivo di *patch* sane e malate.

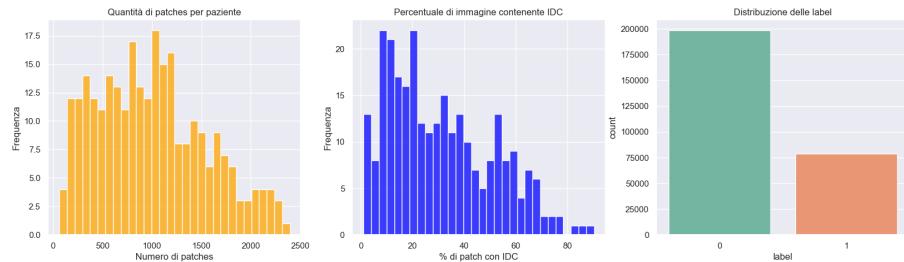


Fig. 1: Distribuzione dataset

Essendo il *dataset* suddiviso per paziente, dove ogni paziente contiene sia *patch* malate che *patch* sane, anche la suddivisione tra *training set* e *test set* viene effettuata per paziente. In particolare si è scelta una suddivisione che prevede il 70% dei pazienti totali per il *training set* e il restante 30 % diviso equamente tra *validation* e *test set*, dato che si prevede la possibilità di dover ottimizzare gli iperparametri di addestramento.

In preparazione al *training*, le immagini sono state normalizzate e ridimensionate in modo da garantire che tutte avessero la stessa dimensione di 50x50 pixel. Inoltre si è effettuata *data augmentation* sul *training set* inserendo dei *flip* orizzontali e verticali.

3 Addestramento dei modelli

3.1 Modello CNN

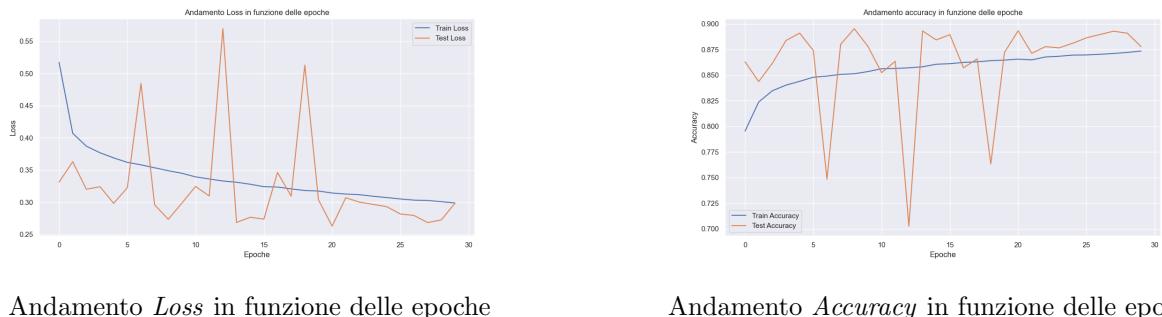
Data la complessità del *dataset*, trattandosi di immagini al microscopio i cui dettagli specifici sono importanti per l'apprendimento, e le sue dimensioni esigue è più indicato usare una rete neurale pre-addestrata ed effettuare un *training* solo su gli ultimi *layers* della rete per adattarla al nostro caso. In quest'ottica si è impiegata la CNN resnet18 a cui si sono aggiunti due layer *fully-connected* con numero di neuroni rispettivamente 512 e 256. Come funzioni di attivazione si sono usate delle ReLU (*Rectified Linear Unit*) in tutti i livelli aggiunti.

Per stabilizzare il processo di apprendimento ed evitare di incorrere nell' *overfitting* si è effettuata *batch normalisation* in ogni *layer* lineare. Si è inoltre aggiunto un *dropout* al 50%.

Il modello è stato poi allenato con *Stochastic Gradient Descent* sul *training set* per 30 epoche con una *batch size* di 32 elementi con *learning rate* impostato a 0.01.

Si sottolinea il fatto che l'obiettivo di questo studio non è l'ottimizzazione delle *performance* del modello CNN, ma il confronto tra due metodologie di addestramento diverse, per cui la scelta di questi parametri è stata ragionevolmente arbitraria, basandosi su valori che vengono categorizzati come valori *standard* nell'ambito del *machine learning*.

In Figura 2 si mostrano i grafici dell'andamento della *loss* e delle *accuracies* di test e addestramento in funzione delle epoche.



Andamento *Loss* in funzione delle epoche

Andamento *Accuracy* in funzione delle epoche

Fig. 2: Risultati dell'addestramento della CNN

3.2 Modello con PCA

A questo punto si è realizzata PCA per verificare se fosse possibile trovare una compressione del *dataset* adeguata per mantenere le stesse prestazioni del modello CNN, ma con tempi di addestramento molto ridotti.

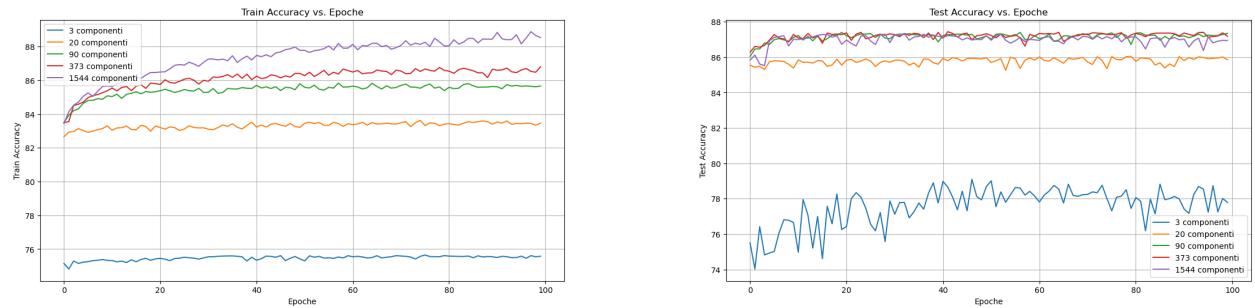
Il parametro fondamentale da decidere per effettuare PCA è il numero di componenti principali che si desidera mantenere.

Per selezionare questo numero si è svolta una PCA che non effettuasse alcuna compressione e si è calcolata l'*explained variance* secondo la formula 1:

$$e.v. = \frac{\sum_{i=1}^d \sigma_i^2}{\sum_{i=D}^D \sigma_i^2} \geq t \quad (1)$$

Per capire come la compressione dei dati influisca sulla capacità di apprendimento del modello, si sono scelti cinque valori diversi di soglia (t) per l'*explained variance*, 0.50, 0.65, 0.75, 0.85 e 0.95 che corrispondono rispettivamente a mantenere le componenti che spiegano il 50%, 65%, 75%, 85% e 95% della varianza totale del *dataset*. A partire da queste considerazioni, si è svolta una PCA per ciascun valore di soglia con le rispettive componenti principali mantenute che sono risultate essere 3, 20, 90, 373 e 1544 a fronte delle 7500 componenti totali del *dataset* originale (50x50x3).

Da ognuna di queste PCA si è ottenuto il *dataset* compresso che è stato poi utilizzato per effettuare l'addestramento del modello. Siccome il *dataset* in seguito alla PCA non è più costituito da immagini a causa della compressione, si può utilizzare un modello non convoluzionale, in particolare si è costruita una semplice rete *shallow* con un solo *layer fully-connected* con 100 neuroni. Si è scelto volutamente un modello semplice in quanto la compressione dovrebbe già aver svolto il lavoro di estrazione dei dettagli importanti per l'apprendimento e un modello più complicato non dovrebbe essere necessario. Il primo addestramento compiuto con i dati compressi ha mostrato segni di *overfitting*: per correggerlo si è effettuata *batch normalization* e si è inserito un *dropout* al 50% come per la CNN. È stato necessario, inoltre, aggiungere una regolarizzazione L_2 , il cui iperparametro *weight decay* è stato stimato facendo 11 addestramenti per individuare il valore migliore, che è risultato essere $\lambda = 0.00005$ (vedi Appendice 6.3). A questo punto è stato possibile realizzare gli addestramenti con i dati compressi; si riportano i risultati nella Figura seguente.



Andamento *train accuracy* in funzione delle epoche

Andamento *test accuracy* in funzione delle epoche

Fig. 3: Risultati degli addestramenti con dati provenienti dalle diverse PCA

3.3 Valutazione dei modelli

Per completare l'analisi si vuole confrontare la *performance* della CNN con la *performance* raggiunta dal modello addestrato con i quattro *dataset* ottenuti dopo la riduzione della PCA.

Per poter confrontare le prestazioni dei modelli in maniera consistente si sono utilizzate diverse metriche, tra cui *accuracy* (2), *precision* (3), *recall* (4) e *F1-score* (5), rispettivamente definite dalle seguenti equazioni:

$$A = \frac{TP + TN}{TOT} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = 2 \frac{P \cdot R}{P + R} \quad (5)$$

dove nel nostro caso i *true positive* (*TP*) sono costituiti dalle *patch* malate che il modello ha predetto essere malate, i *true negative* (*TN*) le *patch* sane che il modello ha predetto essere sane, i *false negative* (*FN*) le *patch* malate che il modello ha classificato come sane e i *false positive* (*FP*) le *patch* sane che il modello ha classificato come malate.

Per meglio visualizzare come variano tali metriche al variare delle componenti principali di ciascuna PCA si è costruito il seguente grafico (Figura 4):

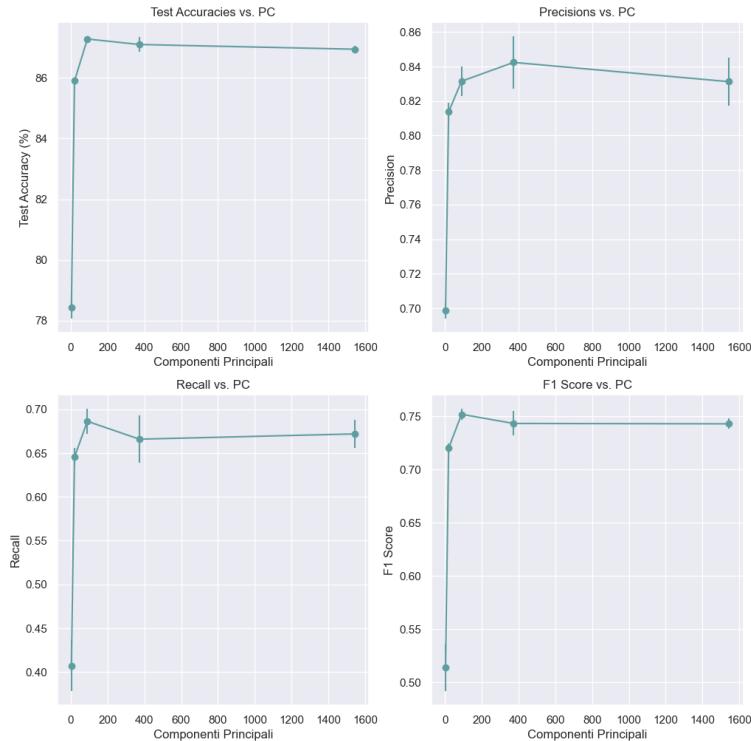


Fig. 4: Andamento delle varie metriche in funzione delle componenti principali.

Per effettuare il confronto con la CNN riportiamo nella seguente tabella i risultati di ciascuna metrica per ciascun modello:

Modello	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F</i> ₁
CNN	0.889 ± 0.004	0.74 ± 0.07	0.82 ± 0.03	0.78 ± 0.05
PCA 3	0.784 ± 0.004	0.698 ± 0.004	0.41 ± 0.02	0.51 ± 0.02
PCA 20	0.8589 ± 0.0009	0.814 ± 0.005	0.65 ± 0.01	0.720 ± 0.004
PCA 90	0.8728 ± 0.0008	0.831 ± 0.008	0.69 ± 0.01	0.751 ± 0.005
PCA 373	0.871 ± 0.002	0.84 ± 0.02	0.67 ± 0.03	0.74 ± 0.01
PCA 1544	0.869 ± 0.001	0.83 ± 0.01	0.67 ± 0.01	0.743 ± 0.005

Table 1: Prestazioni dei modelli

Gli errori associati alle metriche della CNN sono state ottenute come semidisersione dei valori di tre addestramenti successivi mentre gli errori associati alle metriche della rete *shallow* sono stati calcolati come deviazione standard sui valori di 5 addestramenti successivi.

4 Conclusioni

Osservando il grafico in Figura 4 si può notare come la riduzione delle componenti principali inizi ad avere un impatto significativo sulle *accuracies* solo per le PCA da 3 e da 20 componenti. Si potrebbe, quindi, ragionevolmente assumere che questo *dataset* possa essere ridotto ad un minimo di 20 componenti prima di perdere informazioni importanti. Confrontando, invece le metriche di CNN e rete *shallow* (Tabella 1) tra di loro si può notare come, escludendo la PCA più riduttiva di 3 componenti, il modello *fully-connected* addestrato sui *dataset* ridotti produce *score F*₁ che, sebbene inferiori alla CNN, sono comunque ragionevolmente buoni trattandosi di un modello fondamentalmente più semplice della CNN. In generale ciò porterebbe a concludere che svolgere PCA su questo *dataset* porti un notevole vantaggio soprattutto in termini di tempo di addestramento del modello (circa 19 ore per addestrare la CNN e circa 20 minuti per addestrare la rete *shallow* con 2.30 GHz di CPU e 16 GB di RAM). Tuttavia le differenze nei valori di *precision* e *recall* portano a fare alcune considerazioni aggiuntive: come si evince dalla Tabella 1 la rete *shallow* mostra generalmente una *precision* più alta della CNN. Questo significa che addestrando il modello con i dati compressi da PCA esso tenderà a commettere meno falsi positivi rispetto alla CNN, che nel caso in esame significa che la rete *shallow* commette meno errori nel classificare una *patch* sana come malata. Considerando il contesto in cui questo lavoro vuole adoperare significa che la CNN traccerebbe una zona di tessuto malato più ampia rispetto alla realtà. Viceversa, osservando il fatto che i punteggi di *recall* sono generalmente più bassi per la rete *shallow* si può dire che essa è più soggetta a predire dei falsi negativi, cosa che nel contesto si traduce in una più alta probabilità di non individuare una zona di tessuto malato. Sommariamente perciò si può dire che nonostante le prestazioni complessive siano buone, la CNN sarebbe comunque più indicata per svolgere il compito di individuare le zone di tessuto malato. Questo risultato è probabilmente dovuto al fatto che durante la compressione alcuni dettagli importanti per l'individuazione delle *patch* malate possono essere stati rimossi, rendendo il modello meno efficace nella loro individuazione.

Non è da sottovalutare il fatto che, nonostante la PCA a 3 componenti sia quella che ha prodotto le *performance* peggiori in rapporto alle altre, l'*accuracy* del modello rimanga comunque considerevolmente sopra al 50%, nonostante, come si può vedere dalle immagini ricostruite in Appendice 6.4, il *dataset* non presenti più

alcun *pattern* riconoscibile.

Verosimilmente ciò è dovuto al fatto che le *patch* malate tendono generalmente ad avere una colorazione più scura rispetto alle *patch* sane (vedi Appendice 6.2), aspetto che viene ancora mantenuto nella riduzione a tre componenti e che può permettere al modello di fare predizioni corrette nonostante le immagini siano poco riconoscibili.

4.1 Possibili *follow-up*

I dati compressi dalle varie PCA sono stati usati per allenare un modello che è fondamentalmente molto più semplice della CNN. Lo stesso confronto si potrebbe svolgere allenando la stessa CNN ma sulle immagini ricostruite a seguito di una decompressione dei dati ottenuti dalla PCA (vedi Appendice 6.4) in modo da lasciare invariata la complessità del modello.

5 Bibliografia

1. Angel Cruz-Roa et al. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks", Proceedings Volume 9041, Medical Imaging 2014: Digital Pathology; 904103 (2014), DOI: 10.1117/12.2043872
2. <https://www.kaggle.com/paultimothymooney/breast-histopathology-images/data>

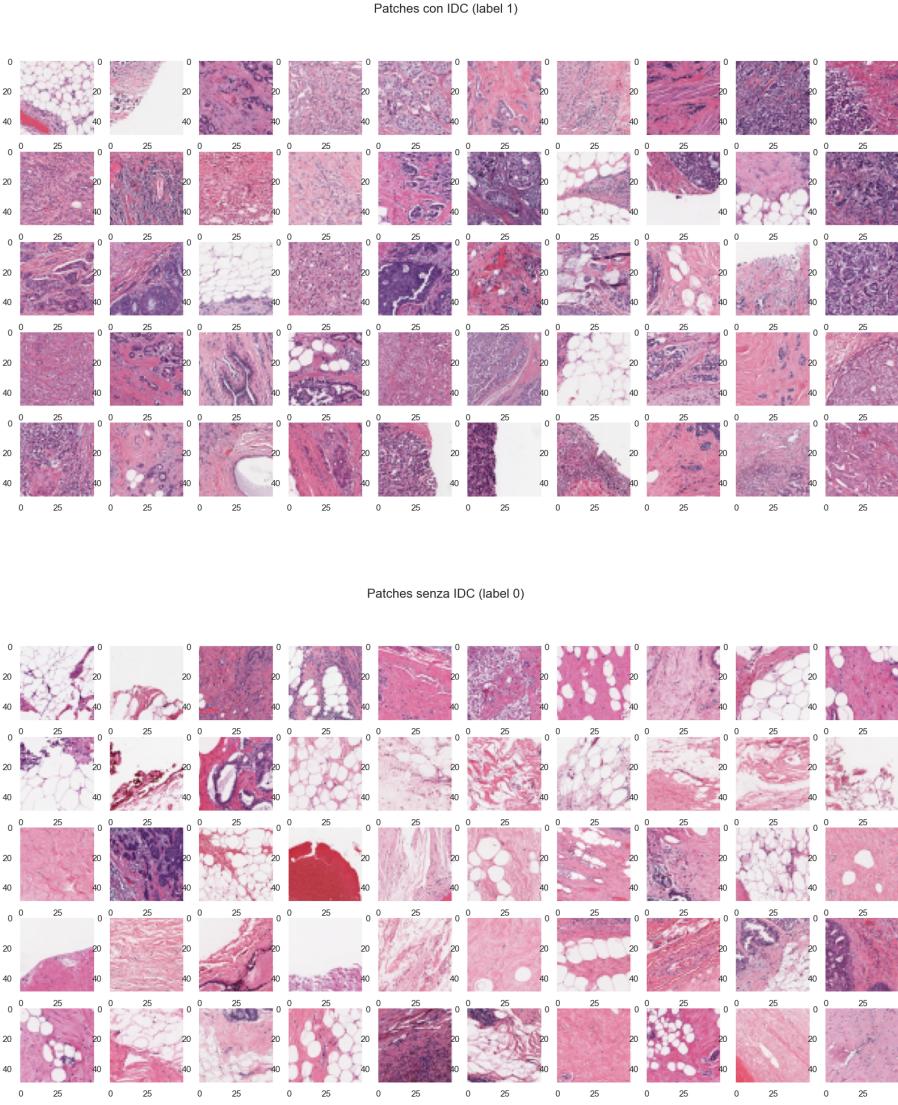
6 Appendice

6.1 Link al codice

<https://github.com/amberymoon/CNN-vs-PCA-in-IDC>

6.2 Immagini dal *dataset*

Si riportano, a titolo di esempio, alcune immagini estratte dal *dataset*.



6.3 Stima del *weight decay*

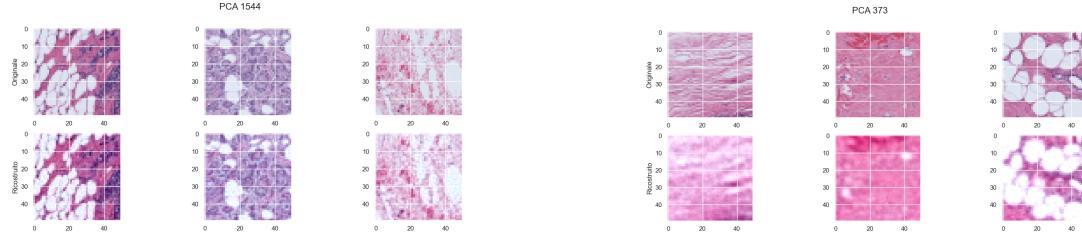
Con l'aggiunta del *weight decay* l'aggiornamento dei parametri con *gradient descent* segue

$$w \leftarrow w - \eta \left(\frac{\partial L}{\partial w} + \lambda w \right)$$

dove η è il *learning rate* e λ l'iperparametro di *weight decay*. Gli addestramenti sono stati svolti sul *validation set* per 11 volte cambiando ogni volta il valore del parametro di *weight decay*. Il parametro selezionato corrisponde a quello che massimizzava la *validation accuracy*.

6.4 Ricostruzioni delle immagini dopo la compressione

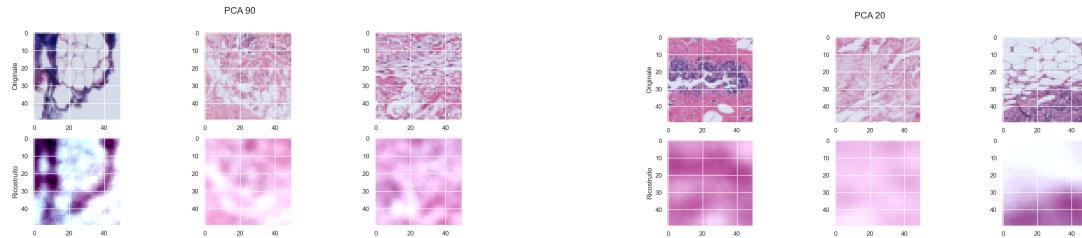
Per avere un'idea di quali dettagli venissero rimossi dalla compressione della PCA si è provato a ricostruire le immagini facendo una decompressione:



Ricostruzione delle immagini per 1544 PC

Ricostruzione delle immagini per 373 PC

Fig. 5



Ricostruzione delle immagini per 90 PC

Ricostruzione delle immagini per 20 PC

Fig. 6

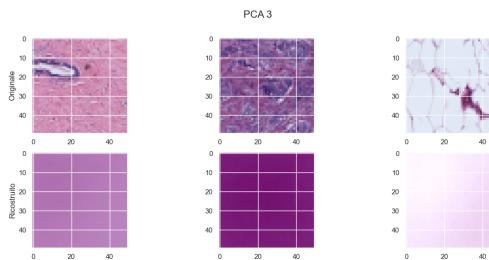


Fig. 7: Ricostruzione delle immagini per 3 PC

Si può notare come i dettagli vengano progressivamente persi man mano che la compressione aumenta. Diventa quindi chiaro il motivo per cui le *accuracies* relative ai dati ottenuti con le PCA da 3 e 20 componenti siano le peggiori in quanto il modello ha pochi dettagli su cui basare il suo apprendimento.