
Università degli Studi di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Corso di Laurea in Informatica

Materia di Tesi: Algoritmi e Strutture Dati

Algoritmi avanzati per la rivelazione di masse tumoriali in mammografia digitale

Tesi di Laurea di:
Massimiliano Zanoni

Relatore:
Chiar.mo Prof. Vittorio Maniezzo

Correlatori:
Chiar.mo Prof. Renato Campanini
Dott. Matteo Roffilli

II Sessione
Anno Accademico 2002-2003

Università degli Studi di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Corso di Laurea in Informatica

Materia di Tesi: Algoritmi e Strutture Dati

Algoritmi avanzati per la rivelazione di masse tumorali in mammografia digitale

Tesi di Laurea di:
Massimiliano Zanoni

Relatore:
Chiar.mo Prof. Vittorio Maniezzo

Correlatori:
Chiar.mo Prof. Renato Campanini
Dott. Matteo Roffilli

Parole Chiave: Mammografia, SVM, Medical Imaging, Wavelet,
Pattern Recognition

II Sessione
Anno Accademico 2002-2003

A mio Padre, a mia Madre e al movimento dei Focolari

Indice

1	Introduzione	1
1.1	Il Tumore al Seno	1
1.1.1	Lo screening di massa	3
1.1.2	Lesioni tumorali	7
1.1.3	Limiti della mammografia	8
1.2	CAD - Computer Aided Detection	12
1.2.1	La mammografia digitale	13
2	SVM e Wavelet	15
2.1	SVM	15
2.1.1	Apprendimento	16
2.1.2	Insiemi separabili	16
2.1.3	Insiemi non separabili	21
2.2	Wavelet	23
2.2.1	Generalità	23
2.2.2	Analisi in Multirisoluzione	24
2.2.3	Haar wavelet	25
2.2.4	Trasformata Wavelet bidimensionale	28
2.2.5	Trasformata Wavelet Overcomplete	31
3	Il CAD	33
3.1	Premesse	33
3.2	Architettura del sistema	34

3.3	Pre-elaborazione	38
3.3.1	Segmentazione	38
3.3.2	Ridimensionamento ed Estrazione dei Crop	39
3.4	Estrazione Feature	42
3.4.1	Trasformata Wavelet	42
3.4.2	Codifica Vettoriale	45
3.5	Addestramento	46
3.6	Riconoscimento	49
3.7	Visualizzazione dell'Output	51
3.7.1	Unione di classificatori	51
3.8	Il CAD parallelo	52
4	Pre-detection	57
4.1	Premesse	57
4.2	Ridimensionamento	60
4.2.1	Convoluzione	62
4.2.2	Filtro Passa Basso Gaussiano	64
4.3	Filtro Passa Alto	67
4.4	Sogliatura (Threshold)	71
4.4.1	Algoritmo di Threshold	73
4.5	Operatori Morfologici	76
4.5.1	Erosione	79
4.5.2	Dilatazione	81
4.5.3	Apertura	82
4.6	Conclusioni	84
5	Risultati	85
5.1	Parametri di valutazione	85
5.2	Database di immagini	86
5.3	Metodologie di valutazione	87
5.3.1	Pre-detection	88
5.3.2	CAD	94

6	Conclusioni e Sviluppi Futuri	99
A	Tabelle - Parametri Grid Searching	101

Prefazione

Il tumore al seno si colloca al primo posto nel mondo in quanto grado di mortalità fra le patologie tumorali che colpiscono la popolazione femminile. Si stima che ogni anno vengano diagnosticati un milione di nuovi casi, con un tasso di mortalità di gran lunga superiore alle trecento mila vittime.

La tecnica di rivelazione del carcinoma mammario attualmente più utilizzato è l'analisi, da parte di un radiologo, della lastra mammografica. Purtroppo l'affidabilità del metodo non è molto elevata: fino al 30% dei casi non vengono diagnosticati se non in seguito a diverse sedute e con il carcinoma già in fase avanzata.

Le cause vanno cercate nella notevole difficoltà della sua individuazione ad occhio nudo, dovuto alla frequente somiglianza con il tessuto ospitante. Ciò accade soprattutto nella fase embrionale dello sviluppo della patologia, in quanto caratterizzata da dimensioni molto ridotte.

Ma proprio la diagnosi precoce che fornisce una possibilità abbastanza elevata di guarigione, garantendo anche una buona possibilità di intervento sul carcinoma con tecniche non invasive.

Si capisce allora l'importanza della diffusione di una politica di prevenzione.

Negli ultimi anni sono andate diffondendosi tecniche di analisi basate sulla diagnosi, dello stesso paziente, da parte di più radiologi e, recentemente, anche sul consiglio di sistemi di rivelazione automatici (Computer Aided Detection - CAD).

La difficoltà di trattamento degli oggetti ricercati fanno di questo un problema

per niente semplice.

Questa tesi vuole essere un contributo alla realizzazione di un sistema CAD, progettato e sviluppato da un gruppo di ricerca dell'Università degli Studi di Bologna, che coinvolge ricercatori del Corso di Laurea in Scienze dell'Informazione (sede di Cesena) e del Dipartimento di Fisica.

Il progetto comprende metodologie per l'elaborazione delle immagini, con applicazioni di tecniche quali le Trasformate Wavelet, allo scopo di porre in evidenza alcune caratteristiche, e sistemi classificatori come la Support Vector Machine (SVM) con il compito decisionale sulla positività dei segnali.

Lo scopo dell'elaborato è quello dello studio di metodologie che migliorino le tecniche esistenti, permettendo di aumentare il numero di lesioni rivelate, diminuendo il numero di falsi positivi prodotti erroneamente dal CAD e abbassando i tempi di calcolo.

In particolare sono state studiate metodologie di segmentazione del mammogramma allo scopo di eliminare preventivamente zone non a rischio, risparmiando l'analisi al classificatore.

Capitolo 1

Introduzione

1.1 Il Tumore al Seno

Il carcinoma mammario è attualmente la forma tumorale più diffusa fra le donne nel mondo. Si stima che nel corso della sua vita, una donna, ha il 12% di probabilità di soffrire del sopracitato disturbo. Inoltre si colloca, in assoluto, al secondo posto come tasso di mortalità femminile.

È sufficiente dare uno sguardo alle statistiche relative all'anno 2000 sotto riportate (*Tabella 1.1*) per accorgersi di come esso sia un problema di portata enorme.

Sono diversi i fattori che, secondo alcuni, possono influire sull'aumento del rischio di sviluppo di un carcinoma mammario: menopausa tardiva, precedenti noduli benigni, terapie ormonali, fattori dietetici, etc.[10] Nel 75% dei casi questi comunque non sembrano essere stati riscontrati. L'unico che è dimostrato assumere un ruolo fondamentale sulla probabilità di contrazione del disturbo è l'età. In uno studio [15] fatto su 1500 donne si è calcolata la distribuzione, sotto riportata, secondo l'età di coloro che hanno sviluppato il tumore al seno. Come si vede il numero di casi aumenta esponenzialmente passata la soglia dei 40 anni, con un picco elevatissimo attorno ai 75 anni

Region	New Cases	Deaths
Africa	59167	19569
Caribbean	621	231
South America	69924	22735
Northern-Central America	220707	57072
Asia	348338	135064
Europe	346118	129698
Australia/New Zealand	12748	3427
Melanesia	470	209
Micronesia	62	28
Polynesia	127	58
Total	1058282	368091

Tabella 1.1: Breast Cancer Statistics-2000 [16]

(Figura 1.1).

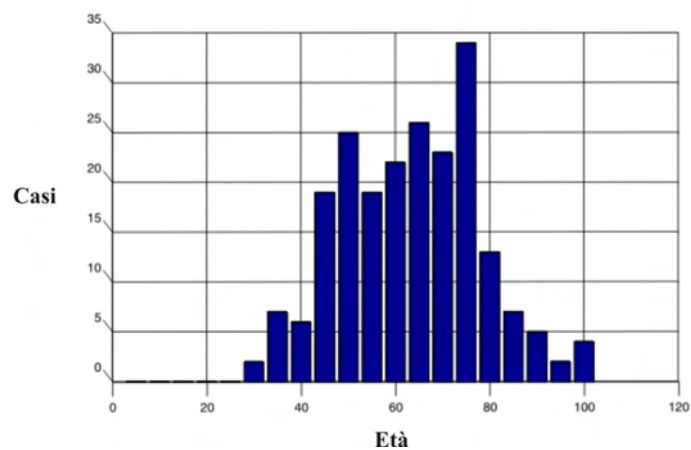


Figura 1.1: Incidenza del carcinoma mammario in funzione dell'età.

Il tumore nasce come patologia locale, limitata alla mammella, con dimensioni, in genere, inferiori a 10mm. Col tempo questo si espande nei tessuti vicini, con la possibilità di contagio di tessuti vitali come i polmoni.

La lotta contro il cancro al seno è fortemente dipendente dal grado di sviluppo del disturbo al momento in cui esso viene diagnosticato: nella malattia scoperta nella sua fase iniziale, è il caso di piccole lesioni, le possibilità di guarigione sono altissime. Si stima che in questi casi la sopravvivenza per i successivi 15 anni è superiore al 90%. Ma la mancanza di segnali sintomatici, se non in fase già avanzate, fanno di questo un compito assolutamente poco semplice. Diventa allora fondamentale adottare una politica di diagnosi precoce con controlli periodici e il più possibile popolari.

In caso di tumore in grado avanzato le terapie adottate sono nella totalità dei casi di demolitivo, come interventi chirurgici per asportazione della mammella seguiti da pesanti trattamenti di radioterapia e chemioterapia, con notevoli conseguenze estetiche e psicologiche. Un tumore ancora allo stadio iniziale, invece, spesso necessita di interventi di tipo conservativo, sola asportazione del nodulo o di una piccola parte della mammella. Ecco quindi che la diagnosi precoce diventa fondamentale anche da un punto di vista terapeutico.

Un grosso passo in avanti è stato fatto anche dalle aziende sanitarie di svariati stati, che mettono a disposizione centri sia di informazione che di controllo, assolutamente gratuiti.

In Italia la campagna di screening di massa in funzione e ogni due anni, le donne tra i 50 e i 60 anni vengono invitate a sottoporsi ad una mammografia. La mammografia per intero a carico del servizio sanitario nazionale. In aggiunta alla gratuità, la particolarità del programma che le strutture sanitarie si fanno totalmente carico delle eventuali malattie riscontrate.

1.1.1 Lo screening di massa

Il metodo più semplice e diretto per esaminare il seno è la palpazione, che può essere eseguita facilmente sia dal paziente che dal medico stesso. Come si può

facilmente intuire però, a causa della natura multi-nodulare della mammella, il metodo non è altamente affidabile. Si riscontra una oggettiva difficoltà nell'individuare i tumori allo stadio iniziale.

Sono state così introdotte altre indagini nella diagnostica senologica, fra queste vi è la *mammografia*.

Questa è una tecnica a raggi-X che consente di avere una visione interna della mammella con una buona risoluzione spaziale e un alto contrasto.

Un fascio di raggi-X, emesso da un apparecchio radiologico detto *mammografo*, attraversa il seno e viene assorbito in dipendenza al tipo di tessuto che incontra. I raggi rimanenti vanno ad impressionare una pellicola. Quello che ne risulta è un'immagine in scale di grigio, rappresentante la struttura interna di una mammella. Le zone radio-opache, che in genere sono quelle più dense come tessuto ghiandolare o fibroso (es. vene), risultano relativamente luminose. Mentre quelle radiolucenti, come il tessuto lipidico, appaiono più scure. Si viene a creare allora una corrispondenza proporzionale fra l'intensità di colore e la densità del tessuto corrispondente.

Le lesioni presenti nel seno hanno un grado di assorbimento caratteristico. Per questo il radiologo nella maggior parte dei casi riesce ad individuarle.

Ciò nonostante, come vedremo poi, vengono commessi ancora molti errori, dovuti alla imperfezione del supporto e del metodo usati.

Uno screening consta, di base, di due proiezioni per ogni mammella: una lungo l'asse *cranio-caudale (CC)* (*Fig 1.2 - Fig 1.3*) ed una *obliqua medio-laterale (MLO)*, inclinata a 45 gradi (*Fig 1.4 - Fig 1.5*).

Al momento dell'esame pieghe temporanee del tessuto dovute al mal posizionamento della mammella o agglomerati venosi, possono apparire come lesioni. In questi casi si richiedono altre analisi di accertamento.

Nel suo complesso l'esame è molto veloce: in genere ha durata che varia fra i 10 e i 15 minuti.

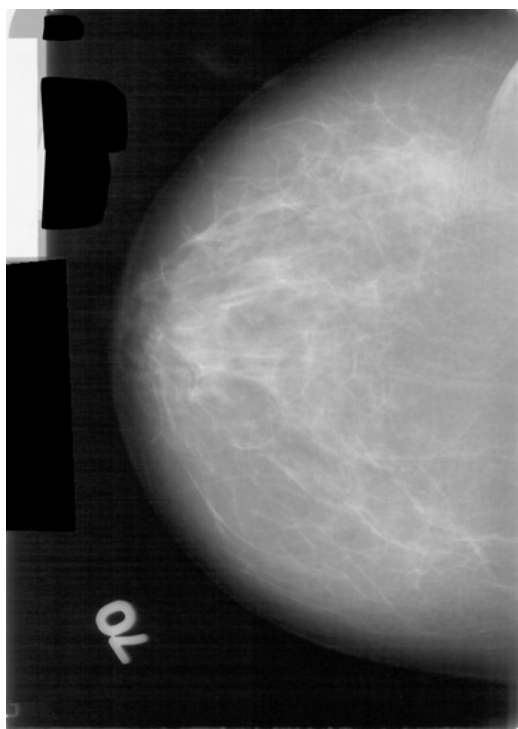


Figura 1.2: Vista cranio-caudale(CC) Sinistra

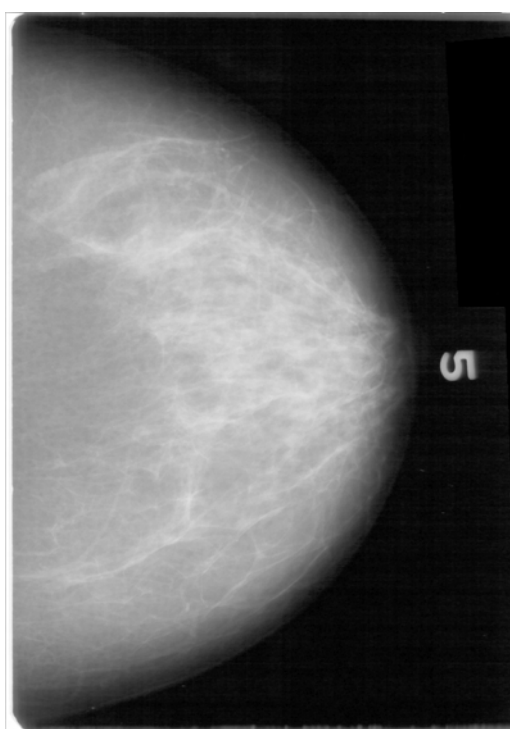


Figura 1.3: Vista cranio-caudale(CC) Destra

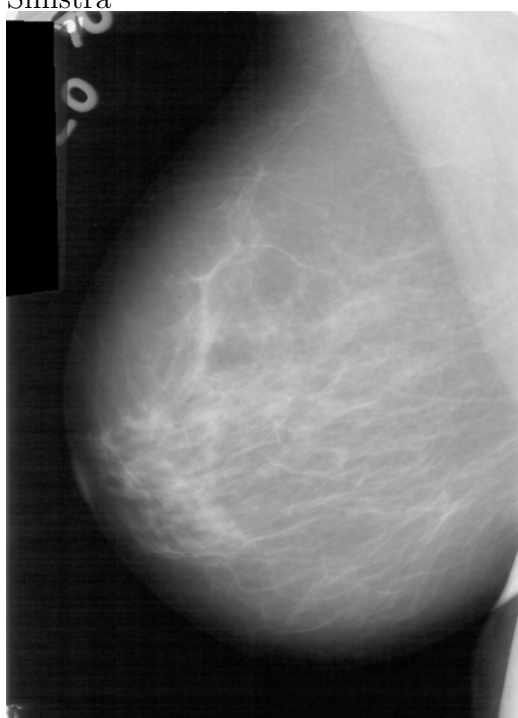


Figura 1.4: Vista obliqua medio-laterale(MLO) Sinistra

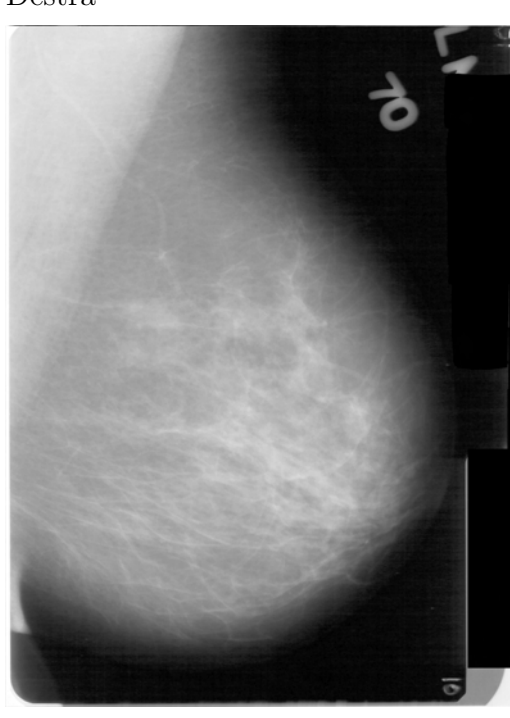


Figura 1.5: Vista obliqua medio-laterale(MLO) Destra

La struttura interna del seno può variare notevolmente da soggetto a soggetto e, di conseguenza, anche il suo aspetto in una mammografia.

Si possono avere mammelle con predominanza di tessuto grasso, che come già detto, è radiolucente e di conseguenza appariranno molto scure e prive di struttura. Oppure principalmente formate da tessuto fibroso che risulteranno molto chiare e con zone ben definite. In questo caso si parla di *mammella densa*(Fig 1.6).

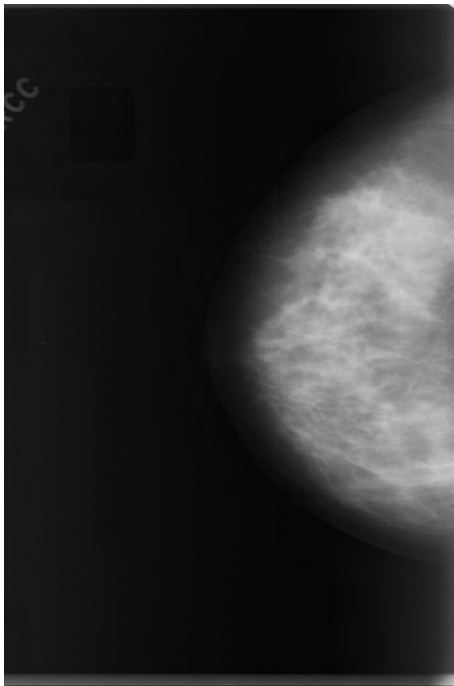


Figura 1.6: Mammella Densa

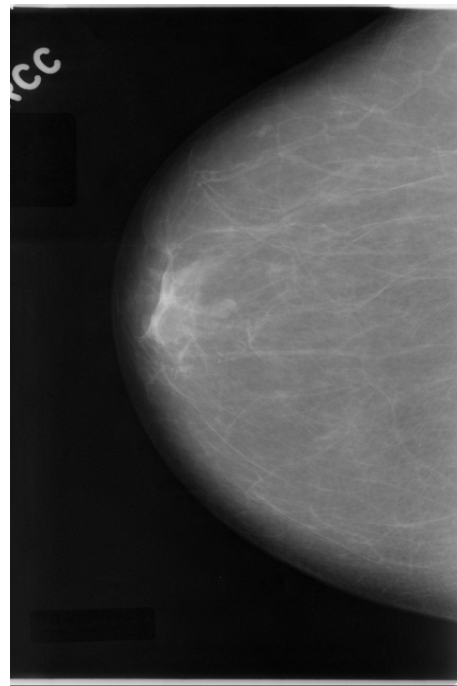


Figura 1.7: Mammella grassa

1.1.2 Lesioni tumorali

Nella mammografia le lesioni tumorali risultano, in genere, come zone molto chiare con caratteristiche particolari, che in un qualche modo le fanno distinguere dal normale tessuto del seno (**parenchimale**). I due tipi più diffusi di patologia sono le **microcalcificazioni** e le **masse tumorali**. Come si può notare in figure 1.8 1.9 esse sono profondamente diversi, necessitano quindi di trattamenti diversi. Nel presente lavoro ci soffermeremo solo sulle masse, tralasciando gli altri tipi di lesione.

È assolutamente difficoltoso fare una classificazione precisa delle masse a causa della loro grandissima variabilità. Possono avere:

- una dimensione fra i 3 e 20-30 mm
- diverse densità ottiche, in base al grado di radiopacità
- bordi più o meno definiti e più o meno regolari.

La loro identificazione deve considerare anche il tipo di tessuto nel quale sono poste. Una massa densa in un tessuto grasso sarà, in genere, facilmente visibile anche ad occhio nudo, in quanto si presenta come una macchia molto luminosa su un contorno scuro. Viceversa sarà difficilmente individuabile una lesione in un tessuto molto denso.

Ciononostante è possibile identificare alcune forme e tipi di bordo classici. Vedi *Tabelle 1.2 e 1.3*.

Data una massa, riuscire a classificarla come combinazione fra forma e tipo di bordo è utile anche per avere informazioni sulla natura stessa della patologia: è stato riscontrato come la morfologia della lesione sia legata in un qualche modo al suo grado di malignità.

Ad esempio masse *spicolate* con nucleo centrale radiopaco (quindi molto luminoso nella mammografia) sono considerate la manifestazione più tipica di

Forma
circolare
ovale
irregolare
distorsione architetturale
asimmetrie di tessuto

Tabella 1.2: Principali forme di masse tumorali

Bordo
circoscritto
oscurato
lobulato
non definito
spicolato

Tabella 1.3: Principali tipi di bordo di masse tumorali

lesioni maligne. Le *spicole* rappresentano la reazione fibrosa dell'organismo ospite alla formazione del tumore.

Due casi particolari sui quali ci soffermiamo brevemente sono le *distorsioni architetturali* e le *asimmetrie di tessuto*. Queste sono diverse dalle altre in quanto non associate a lesioni fisiche visibili. Le prime ad esempio sono zone dove la normale architettura della mammella è distorta, ma non appare nessuna massa tumorale.

Le seconde invece sono asimmetrie che vengono riscontrate fra la mammella destra e quella sinistra, che talvolta possono essere considerate come masse.

Di seguito sono riportati alcuni esempi di lesioni tipiche (*Fig 1.11-1.12-1.13-1.14*).

1.1.3 Limiti della mammografia

La mammografia è l'esame più importante ed utilizzato per la diagnosi del carcinoma della mammella. Tuttavia, nonostante la ricerca abbia migliorato notevolmente la metodologia nel corso negli anni, ha ancora alcuni problemi e si stima che sulla totalità delle lesioni neoplastiche mammarie non vengano diagnosticati fra il 10% e il 30% dei casi.

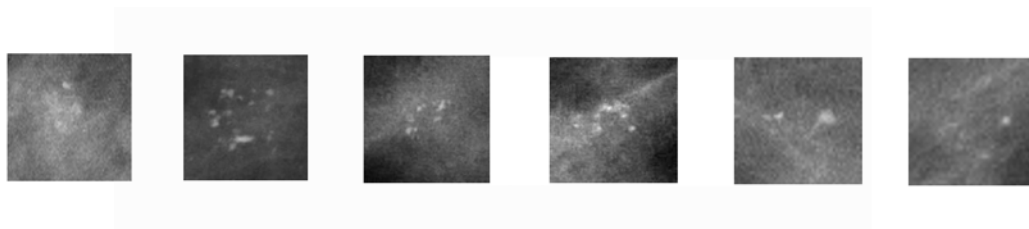


Figura 1.8: Microcalcificazioni



Figura 1.9: Masse Tumorali

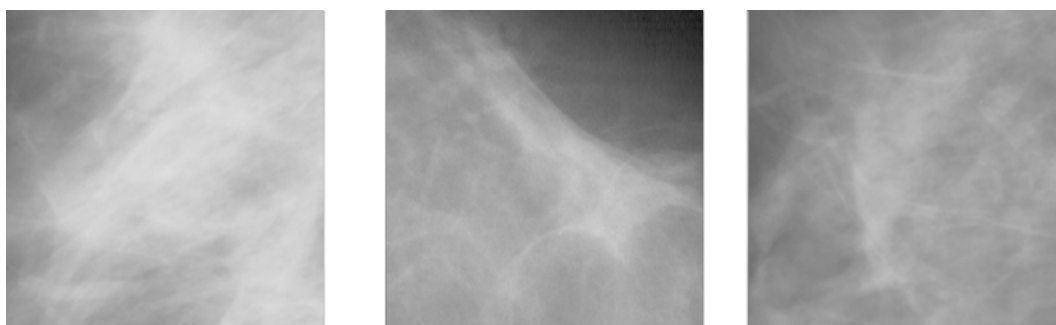


Figura 1.10: Tessuti Sani

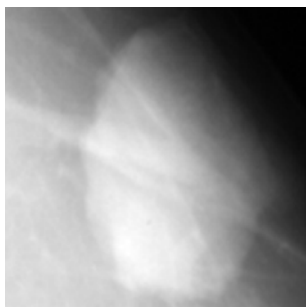


Figura 1.11: Massa Lobulata

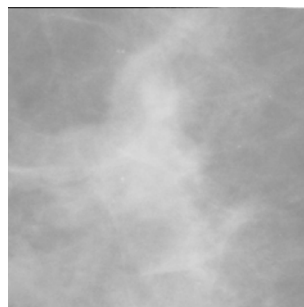


Figura 1.12: Massa non definita



Figura 1.13: Massa Spicolata

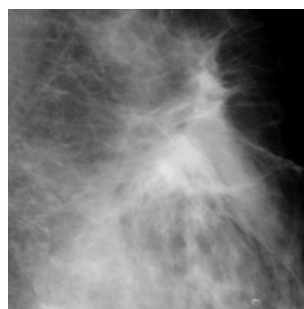


Figura 1.14: Distorsione Architetturale

Seguendo i passi successivi dei quali è composta l'attuale metodologia di analisi si può riscontrare come siano molteplici i fattori che contribuiscono ad una percentuale di errore così elevata.

Primo fra tutti la natura stessa della massa. Si è già accennato alla grande variabilità delle sue caratteristiche morfologiche, e quanto la sua individuazione sia connessa al contrasto intrinseco esistente con tessuto ospitante. Grossi problemi si hanno in genere in mammelle di pazienti giovani, nelle quali si presenta un tessuto fibroso e quindi una mammella molto densa.

Tutto il processo di acquisizione delle informazioni e loro stampa sul supporto finale aggiunge rumore alle informazioni di partenza. Per cui il radiologo, andando ad analizzare una mammografia affetta da rumore di fondo, avrà più probabilità di commettere errori.

L'ultimo stadio del processo è il verdetto del radiologo stesso. Qui, oltre a quelli

sopraindicati, subentrano i fattori di imperfezione umana. Due frequenti cause di errore sono:

- l'affaticamento degli occhi dovuto a lavoro troppo intenso
- la mancanza di tempo, per cui inconsciamente si analizzano solo le parti dell'immagine nelle quali pensiamo sia più probabile trovare masse

Ricerche di Psicofisica¹ hanno dimostrato quanto questi fattori siano condizionanti la buona riuscita della diagnosi [11] [23].

Si distinguono due tipi di errori:

- falso negativo (*false-negative error*), che si verifica quando una mammografia contenente un qualche tipo di lesione viene erroneamente classificata come normale
- falso positivo (*false-positive error*) che viene commesso quando in una normale mammografia vengono segnalate lesioni che, invece, non esistono.

L'errore chiaramente più grave quello di tipo falso-negativo, in quanto provoca un ritardo nella diagnosi e nella cura del disturbo, che potrebbe compromettere irrimediabilmente la salute della paziente. Un errore di tipo falso-positivo, sebbene non comprometta la probabilità di sopravvivenza della paziente porta a conseguenze psicologiche non trascurabili. In particolare è stato provato che il livello di ansietà in donne richiamate per ulteriori accertamenti risulta molto più elevato rispetto a quello di donne che si presentano allo screening mammografico per un normale controllo [24].

¹Scienza che si occupa del rapporto fra il mondo fisico ed esperienza mentale dello stesso.

1.2 CAD - Computer Aided Detection

Data l'importanza di una diagnosi precoce e in, contrapposizione, le difficoltà riscontrate col tempo affiancati al radiologo meccanismi che lo possano in qualche modo aiutare ad abbassare le possibilità di errore.

All'inizio degli anni 70 è stato introdotto l'uso del computer nella mammografia per il miglioramento dell'immagine: aumento del contrasto, filtri...

Poi sono state sviluppate tecniche di analisi nelle quali venivano impiegati due radiologi (doppia analisi): entrambi facevano la loro diagnosi indipendentemente e se ne discutevano poi i risultati. Il successo del metodo (diagnosi corretta nell'85% dei casi [23]) ha portato ad una sua evoluzione: la sostituzione di uno dei due umani con un sistema di rilevamento automatico detto appunto CAD.

È importante sottolineare che la sentenza del CAD non è da intendersi esaustiva, ma solo indicativa al radiologo delle regioni sospette.

Generalmente un CAD consiste di due fasi. La prima detta **detection** ha lo scopo di determinare le zone dell'immagine, che per le loro caratteristiche potrebbero ospitare una lesione, dette **ROI - Region of Interest**. La seconda è quella di **classificazione** nella quale il sistema è in grado di decidere se ognuna di queste zone è o meno una massa tumorale ed eventualmente se è benigna o maligna.

L'efficienza di un sistema CAD vengono valutato secondo i seguenti parametri:

- Percentuale di **veri positivi (true positive)** individuati
- Numero di Falsi Positivi - Con un numero di falsi positivi troppo alto ($>2-3$) il sistema sarebbe inutile, il radiologo infatti non ne troverebbe vantaggio in quanto dovrebbe discriminare fra questi segnali.
- Tempo di elaborazione

I sopracitati saranno i parametri sui quali ci baseremo per valutare il CAD del presente lavoro.

1.2.1 La mammografia digitale

L'utilizzazione di un sistema CAD in appoggio al radiologo necessita che la pellicola mammografica impressa dal mammografo sia portata in digitale per permettere poi l'elaborazione su di un calcolatore. Per diminuire il più possibile la perdita di informazioni il processo viene eseguito attraverso scanner ad altissima definizione. Ma, comunque, qualsiasi discretizzazione porta all'aggiunta di rumore.

Se già da diversi anni in molti capi questo problema è stato ampiamente superato prelevando le informazioni direttamente in digitale, nella mammografia questo è stato possibile solo ultimamente. Il ritardo è dovuto principalmente alla complessità e di conseguenza dei costi delle apparecchiature richieste.

Dopo oltre dieci anni di ricerche si è arrivati alla realizzazione, ad un costo accettabile, del **mammografo digitale**. La pellicola radiografica è sostituita da un rivelatore digitale che converte la quantità di raggi-X attraversanti la mammella in segnali elettrici, i quali vengono memorizzati come segnali digitali. Questi andranno a comporre un mammogramma digitale che visualizzabile su di uno schermo o stampato su carta.

Rispetto alle tecniche classiche i vantaggi sono chiaramente innumerevoli. I primi due la possibilità di modificare il mammogramma in seguito al prelievo delle informazioni. La pellicola è un supporto non modificabile, mentre una immagine memorizzata sul computer sì. Questo permette di applicare fasi di image processing, quali filtri e aumenti del contrasto e luminosità, per migliorare l'immagine, in maniera assolutamente automatica.

Un altro beneficio apportato è che, idealmente, non ci sono vincoli alla risoluzione spaziale dell'immagine, per cui in generale si ha una qualità diagnostica nettamente migliore. Se ne porta un esempio in figura 1.16. Qui si vede il confronto fra le tecnologie: in quella digitale vi è quasi assenza di rumore con conseguente nitidezza e precisione sui dettagli molto più elevata.

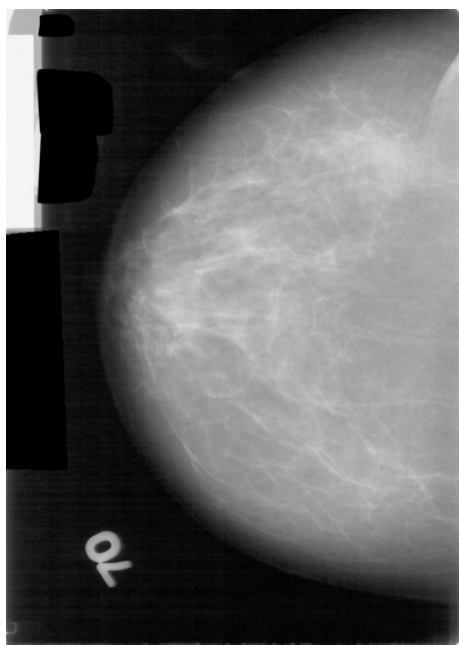


Figura 1.15: Mammogramma digitalizzato

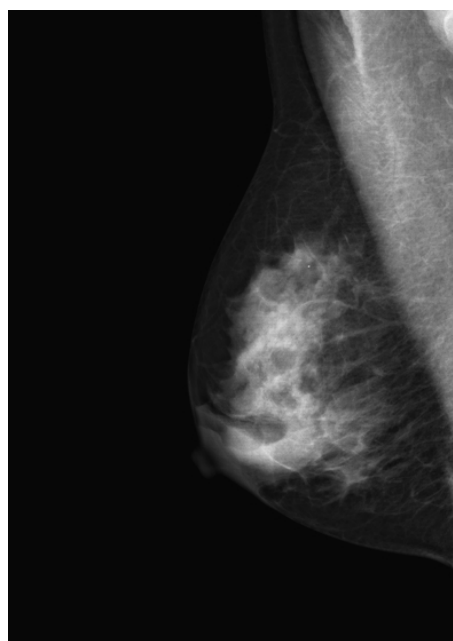


Figura 1.16: Mammogramma digitale

In conclusione, il trattamento di tutti i dati di un paziente in formato digitale dà la possibilità in futuro di avere la propria cartella clinica non più cartacea, ma semplicemente come una directory su una macchina in qualche luogo. Non è nemmeno più importante in quale clinica sia sita. Al momento del bisogno è trasferibile in pochi minuti da un capo all'altro del mondo. È superfluo evidenziare quali siano gli enormi vantaggi in termini di tempo e costo.

Capitolo 2

SVM e Wavelet

2.1 SVM

Il compito della fase di classificazione consiste nel discriminare fra lesioni tumorali e zone sane, sulla base di alcune informazioni (caratteristiche) che vengono fornite.

Tale problema, può essere concettualmente ricondotto alle classi di equivalenza: si vuole individuare la funzione di proiezione tale che, un elemento del dominio di riferimento venga giustamente e univocamente mappato sulla sua classe di appartenenza.

Si definisce (Gonzales *at al.* [13]) **Pattern Recognition** come:

Un processo attraverso il quale un elemento è assegnato ad una fra un prescritto numero di classi

Nel quadro della ricerca scientifica sta ultimamente riscuotendo molti apprezzamenti la tecnica delle **Support Vector Machine** (SVM)[2][9].

2.1.1 Apprendimento

Il termine apprendimento fa pensare ad una risposta adattiva da parte del sistema sulla base di stimoli di feed-back provenienti dal mondo esterno, che premiano o penalizzano una la risposta data. Riportato al *pattern recognition*, il problema può risultare relativamente semplice se dato un valore di input conosciamo già la sua classe di appartenenza.

Non è il nostro caso.

L'SVM si colloca fra le tecniche di apprendimento statistico, le quali trovano applicazione proprio nei casi in cui non sia possibile conoscere la funzione di proiezione fra l'oggetto da classificare e la sua classe di appartenenza. Il sistema deve quindi cercare di dedurre la relazione funzionale, detta *funzione obbiettivo*, attraverso un set di esempi. Chiaramente fra tutte le possibili si considera solo la classe (*spazio delle ipotesi*) di quelle aventi come dominio il nostro insieme degli elementi da classificare, e come codominio l'insieme delle classi di arrivo. L'euristica ottenuta della *funzione obbiettivo* viene definita *funzione decisionale*.

L'SVM è un *algoritmo di apprendimento* che dato un insieme di esempi e ed uno *spazio delle ipotesi* determina una *funzione decisionale* più o meno buona.

2.1.2 Insiemi separabili

Affrontiamo il problema del *Pattern Recognition* da un punto di vista geometrico.

Supponiamo di avere l elementi, come rappresentati in Figura 2.1, ognuno dato dalla coppia:

$$(\mathbf{x}_i, y_i) \quad \text{per} \quad i = 1, \dots, l$$

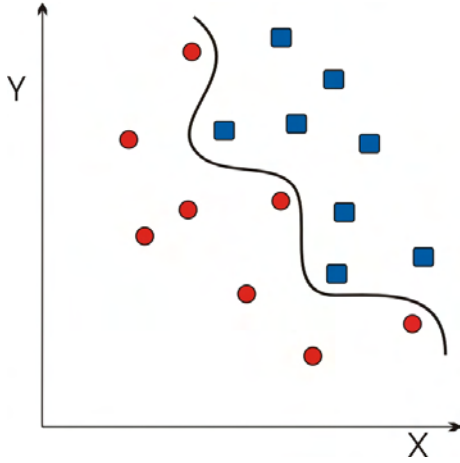


Figura 2.1: Superficie di separazione delle classi

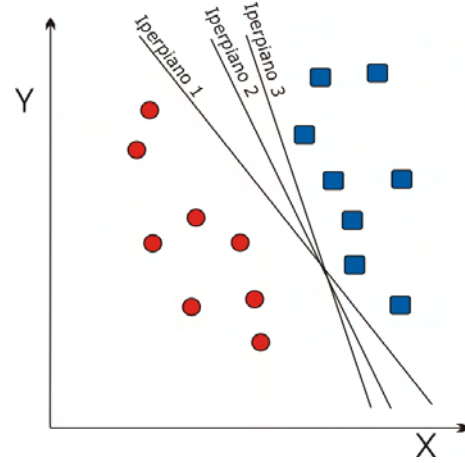


Figura 2.2: Classificazione binaria con funzioni lineari

$\mathbf{x}_i \in \mathbf{X} \subseteq \mathbf{R}^n$ vettore per l' i -simo elemento

$y_i \in \{-1, +1\}$ valore di appartenenza ad una classe

La classificazione consiste nella ricerca di una superficie che separi le due classi di elementi.

Se ci restringe al caso in cui la superficie suddetta sia un iperpiano, si parla di *classificatore lineare* (Figura 2.2).

La funzione separatrice sarà del tipo $f : \mathbf{X} \subseteq \mathbf{R}^n \rightarrow \mathbf{R}$ definita come:

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

essendo questa la funzione di un iperpiano¹.

Da qui possiamo ricavare che un iperpiano separatore è definito come:

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$$

dove w è la normale all'iperpiano e $\frac{b}{\|\mathbf{w}\|}$ è la sua distanza dall'origine².

¹Il prodotto scalare è inteso essere quello euclideo.

²La norma è intesa essere quella euclidea

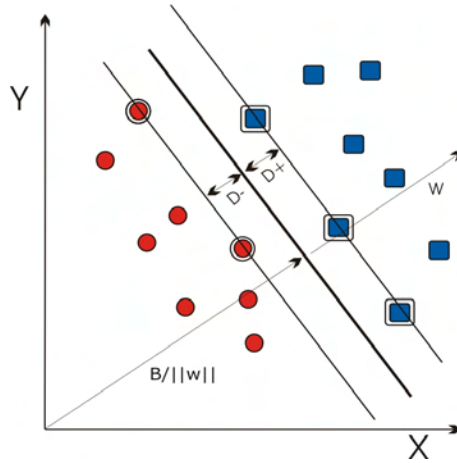


Figura 2.3: Piano di separazione lineare. Gli elementi cerchiati sono i support vector

Come è intuibile (Figura 2.2) esistono infiniti iperpiani di separazione; ci poniamo quindi il problema della scelta dei parametri per la discriminazione di uno rispetto ad un altro nella ricerca dell'ottimale.

Consideriamo i due iperpiani di confine delle classi (Figura 2.3). Siano d_+ e d_- rispettivamente le distanze fra il piano ottimo e i più vicini positivi ($y_i = 1$) e negativi ($y_i = -1$). Definiamo *margin* come $\Delta = d_+ + d_-$.

L'SVM si basa sulla tecnica detta *Maximal Margin Hyperplane* che consiste nel massimizzare la distanza Δ . Il principio fondante è che in prossimità di un punto appartenente ad una certa classe, è più probabile trovare punti della stessa classe che punti dell'altra. Pertanto massimizzando la distanza delle classi aumenta la probabilità di corretta generalizzazione.

Si può dimostrare che dato un insieme di esempi di training

$$(\mathbf{x}_i, y_i) \quad \text{per} \quad i = 1, \dots, l$$

esso è linearmente separabile se esiste almeno un \mathbf{w} tale che vale la seguente relazione:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \quad (2.1)$$

La 2.1 si può spezzare in due relazioni che evidenzino ciascuna classe:

$$(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq +1 \quad \text{per} \quad y_i = +1 \quad (2.2)$$

$$(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq -1 \quad \text{per} \quad y_i = -1 \quad (2.3)$$

Dalle relazioni 2.2 e 2.3 si ricava che i punti residenti sui piani di confine sono caratterizzati da $(\langle \mathbf{w}, \mathbf{x} \rangle + b) = +1$ e $(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq -1$ con distanze dall'origine di rispettivamente $\frac{|1-b|}{\|\mathbf{w}\|}$ e $\frac{|-1-b|}{\|\mathbf{w}\|}$. Da cui $d_+ = d_- = \frac{1}{\|\mathbf{w}\|}$ e dunque $\Delta = \frac{2}{\|\mathbf{w}\|}$.

Allora il problema di MMH si riduce a:

$$\text{minimizzare} \quad \|\mathbf{w}\|^2 \quad (2.4)$$

sotto le condizioni della 2.1

Quello formulato rientra nella classe di problemi di Ottimizzazione Convessa Quadratica (Convex Quadratic Program): funzione da minimizzare quadratica e vincoli lineari.

Ora viene fatto un passaggio alla rappresentazione tramite moltiplicatori di Lagrange che ci permetteranno una risoluzione più semplice.

Dati $\alpha_i, i = 1, \dots, l$ moltiplicatori di Lagrange positivi, definiamo il Lagrangiano come:

$$L_p(\mathbf{w}, b, \alpha) = \|\mathbf{w}\|^2 - \sum_i \alpha_i y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_i \alpha_i \quad (2.5)$$

Quello che si vuole è allora, sotto le condizioni della 2.1, trovare un minimo delle L_p . In questo ci viene in aiuto il teorema di Karush-Kuhn-Tucker (KKT) [19] secondo il quale si ricava che la formulazione del suddetto problema è equivalente a trovare una coppia $(\|\mathbf{w}\|, \alpha)$ ed uno scalare b tale che valgono le seguenti proprietà:

$$\frac{\partial}{\partial \mathbf{w}} L_p(\mathbf{w}, b, \alpha) = 0 \quad (2.6)$$

$$\frac{\partial}{\partial b} L_p(\mathbf{w}, b, \alpha) = 0 \quad (2.7)$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \quad (2.8)$$

$$\alpha_i[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] = 0, \alpha_i \geq 0 \quad (2.9)$$

Notare che richiedere che il gradiente si annulli, significa cercare un punto di massimo o di minimo.

Si ottiene:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad \text{con} \quad \sum_i \alpha_i y_i = 0$$

Può risultare più conveniente rappresentare il Lagrangiano in termini del suo Duale (Wolf dual[19]):

$$L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \mathbf{x}_j \rangle \quad (2.10)$$

$$\text{con } 0 \leq \alpha_i \text{ e } \sum_i \alpha_i y_i = 0$$

Il problema allora si riduce nel massimizzare la 2.10.

Osserviamo che $\alpha_i \neq 0$ sono nel caso in cui $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 = 0$, cioè quando il punto \mathbf{x}_i appartiene ad uno dei due iperpiani paralleli a quello ottimo, descritti in precedenza. Questi punti sono detti **Support Vector** ed è solo su di essi che viene eseguita la 2.10. Intuitivamente, per massimizzare la distanza Δ ci interessano sono i punti più vicini.

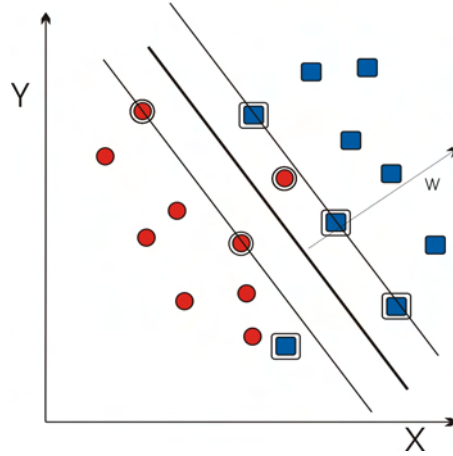


Figura 2.4: Insieme di training non linearmente separabile

2.1.3 Insiemi non separabili

Nel caso di un set di dati di training non linearmente separabile, non è possibile applicare le relazioni viste nel paragrafo precedente.

Quello che accade è nella classificazione, la SVM, commetterà degli errori, a causa della natura stessa del problema. Viene così introdotto un nuovo vincolo: si vuole massimizzare la distanza fra gli iperpiani al confine delle classi e minimizzare il numero totale di errore. Per fare ciò è necessario formalizzare il concetto di errore e, viene fatto attraverso l'introduzione di variabili dette *slack variable* ξ_i .

I vincoli dovranno allora:

$$(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq +1 - \xi_i \quad \text{per} \quad y_i = +1 \quad (2.11)$$

$$(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq -1 + \xi_i \quad \text{per} \quad y_i = -1 \quad (2.12)$$

Si consideri un caso errato nel quale viene classificato positivo ($y_i = +1$) ciò che positivo non è. Si avrà $y_i = +1$ e $\langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1$ e, quindi, per la 2.8 implica $\xi_i > 1$.

Allora $\sum_i \xi_i > 1$ è un limite superiore al numero di errori sull'insieme di train-

ing.

Il problema può essere allora riformulato come:

$$\text{minimizzare} \quad \|\mathbf{w}\|^2 + C\left(\sum_i \xi_i\right)^k \quad (2.13)$$

sotto le condizioni 2.11 e 2.12, dove C è un reale positivo che definisce il peso da attribuire agli errori e k un intero positivo che definisce il grado.

Si nota come questo sia ancora un problema Convesso $\forall k \in \mathbf{N}$ ed in particolare un Convex Quadratic Problem per $k = 1, 2$. Il caso più vantaggioso si ha per $k = 1$ per cui il duale di Wolf, che senza entrare troppo in formalismi, assume la forma:

$$L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \mathbf{x}_j \rangle \quad (2.14)$$

con $0 \leq \alpha_i \leq C$ e $\sum_i \alpha_i y_i = 0$

L'unica differenza dalla 2.10 è per il limite superiore ad α_i , C , detto *parametro di regolarizzazione*.

Allora una volta che il la SVM è stato addestrato, cioè è stato risolto il Convex Quadratic Problem di un set di esempi, la classificazione di un nuovo vettore \mathbf{x} , Maximal Margin Hyperplane Problem, è dato dalla soluzione della seguente relazione:

$$f(x) = \text{sign}\left(\sum_i y_i \alpha_i (\langle \mathbf{x}, \mathbf{x}_i \rangle) + b\right) \quad (2.15)$$

dove gli \mathbf{x}_i sono i soli vettori di supporto e b è dato da:

$$b = \frac{1}{2} \left[\langle \mathbf{w}, \mathbf{x}_i^{(1)} \rangle + \langle \mathbf{w}, \mathbf{x}_i^{(-1)} \rangle \right] \quad (2.16)$$

con $\mathbf{x}_i^{(1)}$ e $\mathbf{x}_i^{(-1)}$ vettori di supporto relativi alle classi $y = 1$ e $y = -1$

Per concludere si vuole accennare a due ulteriori sviluppi del sistema formale sopraindicato, i quali non è scopo di codesto lavoro analizzare:

- Nella rappresentazione data, si è attribuito lo stesso peso per errori di entrambe le classi. Si può pensare ad un formalismo in cui si discrimini il tipo di errore commesso con l'assegnazione di pesi C differenti.
- È stato analizzato solo il caso in cui la superficie di separazione sia lineare. Nessuno vieta di considerarne altre con gradi superiori. Lo svantaggio derivante è chiaramente l'aumento della complessità.

2.2 Wavelet

2.2.1 Generalità

Una immagine digitale, nel pensiero comune, è un insieme di elementi raggruppati secondo un ordine prestabilito e rigoroso caratterizzati da uno o più valori di luminosità che ne rappresentano il colore. Una immagine in bianco e nero, ad esempio, è una matrice di numeri interi ripartiti secondo un prestabilito intervallo di valori i cui estremi sono il bianco e il nero.

In genere questi elementi vengono sistemati in strutture bidimensionali o tridimensionali. Le grandezze di tali dimensioni sono dette *risoluzioni spaziali*.

Ritornando all'esempio, si può parlare di immagine in scala di grigi a 250 valori, di dimensioni 200x200 pixel.

Quella appena data non è altro che la *rappresentazione* dell'immagine. Cioè il modo di rappresentare, raggruppare gli elementi. Quella matriciale a livelli di grigio ne è il più comune.

Definiamo *trasformazione* un processo di codifica dei dati, identificato da relazioni matematiche, che permetta di cambiare la rappresentazione dell'immagine.

La trasformazione non comporta perdita d'informazione se la relazione indotta fra l'oggetto in input e quello in output è biunivoca.

Gli scopi per cui si voglia o si debba far uso di trasformate sono fondamentalmente due:

- *Compressione* - Essendo digitale, perchè l'immagine esista, è necessario che i dati vengano memorizzati. Può essere abbastanza intuitivo che un modo di rappresentarli può essere più o meno compatto di un altro.
- *Mettere in evidenza particolari* - Ci si è accorti che alcune codifiche possono essere più informative di altre, relativamente all'obiettivo cercato.

La trasformata più famosa è sicuramente la Trasformata di Fourier per la quale un segnale (ad esempio un'immagine) può sempre essere vista come composizione di funzioni sinusoidali di supporto infinito ³ di frequenza variabile. Le funzioni di base della Trasformata Wavelet, invece, sono a supporto compatto, ovvero sono diverse da zero solo su una porzione del suo dominio di definizione, con l'implicazione che oltre a variare in frequenza lo fanno anche nello spazio. Riescono quindi a caratterizzare anche informazioni riguardo lo spazio di applicazione. Questo è estremamente utile per porre in evidenza sia proprietà che riguardano le frequenze, sia la struttura relativa a zone del dominio.

Nello specifico dell'Elaborazione delle Immagini, le trasformate Wavelet sono molto utili per mettere in evidenza alcune caratteristiche strutturali di zone dell'immagine.

2.2.2 Analisi in Multirisoluzione

La teoria di base delle trasformate Wavelet si basa sul concetto di **Analisi in Multirisoluzione**. L'idea è che a diverse scale, possono essere posti in evidenza particolari differenti dell'immagine. Questo è visto molto bene nella cartografia: se vogliamo analizzare la forma dei continenti faremo uso di una

³Per supporto si intende la parte del dominio nel quale la funzione è diversa da zero.

cartina planetaria, quindi con scala molto grande; viceversa se ciò che ci interessa sono le vie di una città ci affideremo ad una mappa con una scala molto più piccola.

La tecnica di analisi in multirisoluzione permette di analizzare l'immagine a diverse risoluzioni, per poterne cogliere diversi tipi di informazioni fornite da ognuna.

Più formalmente possiamo definirla nel seguente modo.

Una possibile rappresentazione di una immagine è di vederla come una funzione costante a tratti sull'intervallo $[0, 1)$. Un'immagine di un solo pixel sarà data da una funzione costante su tutto l'intervallo $[0, 1)$. Si chiami V^0 lo spazio vettoriale di tutte tali funzioni. Un'immagine con due pixel sarà data da una funzione costante sugli intervalli $[0, \frac{1}{2})$ e $[\frac{1}{2}, 1)$. V^1 sarà lo spazio vettoriale di tutte tali funzioni. Lo spazio V^j sarà composto da tutte le funzioni costanti sui 2^j intervalli di divisione di $[0, 1)$. Come si può intuire la successione degli spazi vettoriali definiti è annidata: $V^0 \subset V^1 \subset V^2 \dots$

La successione è proprio detta Analisi in Multirisoluzione.

2.2.3 Haar wavelet

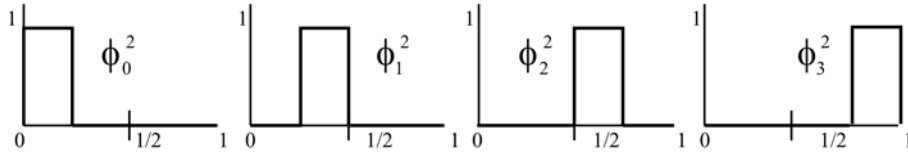
Si definisca una base di funzioni per gli spazi vettoriali sopracitati come:

$$\phi_i^j(x) = \sqrt{2^j} \phi(2^j x - i) \quad i = 0, \dots, 2^j - 1 \quad (2.17)$$

Le funzioni $\phi(x)$, dette *funzioni di scala* (*scaling functions*), sono definite come:

$$\phi(x) = \begin{cases} 1 & \text{per } 0 \leq x < 1 \\ 0 & \text{altrimenti} \end{cases} \quad (2.18)$$

Il passo successivo è quello di definire prodotto fra gli elementi di uno spazio:

Figura 2.5: Funzioni di base per lo spazio V^2

$$\langle f, g \rangle = \int_0^1 f(x)g(x)dx \quad (2.19)$$

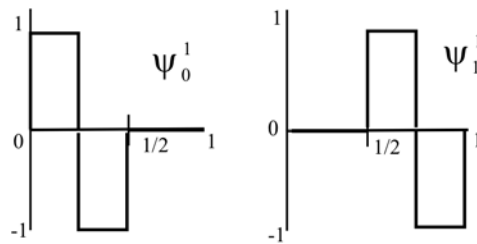
Si ponga ora W^j , detto *sottospazio wavelet*, il complemento ortogonale di V^j e V^{j+1} . Cioè lo spazio delle funzioni $f(x)$ e $g(x)$ ortogonali fra loro secondo il prodotto sopra definito.

Un base per W^j sono le *funzioni Haar wavelet* definite come:

$$\psi_k^j(x) = \sqrt{2^j} \psi(2^j x - k) \quad k = 0, \dots, 2^j - 1 \quad (2.20)$$

dove

$$\psi(x) = \phi(x) = \begin{cases} 1 & \text{per } 0 \leq x < \frac{1}{2} \\ -1 & \text{per } \frac{1}{2} \leq x < 1 \\ 0 & \text{altrimenti} \end{cases} \quad (2.21)$$

Figura 2.6: Funzioni Haar per lo spazio W^1

Gli insiemi di funzioni definiti hanno delle importanti proprietà:

- Le funzioni base $\phi_i^j(x)$ di V^j e le funzioni base $\psi_i^j(x)$ di W^j formano una base di V^{j+1}
- Ogni funzione $\phi_i^j(x)$ è ortogonale alle corrispondente $\psi_i^j(x)$

Dalla prima proprietà si può dimostrare che ogni funzione appartenente a V^{j+1} è sempre esprimibile come combinazione fra una funzione di V^j e una di W^j . Rapportato all'elaborazione delle immagini, questo dice che una immagine è sempre ricostruibile da una con risoluzione minore, dati alcuni coefficienti aggiuntivi forniti dalle Wavelet. In effetti la funzioni Wavelet non sono altro che una rappresentazione delle informazioni che si sono perse durante il processo del cambiamento della risoluzione.

Il processo eseguito ricorsivamente prende il nome di *filter bank*.

Facciamo un esempio:

si consideri una immagine mono-dimensionale con risoluzione 4 pixel:

[8 4 1 3]

Ora viene costruita una nuova immagine, con risoluzione più bassa, facendo la media dei pixel vicini: [6 2]

Come si vede è avvenuta una perdita di informazioni in quanto non saremmo più in grado di eseguire il processo inverso. È necessario aver dei dati aggiuntivi, detti coefficienti di dettaglio. In questo caso vengono scelti [2 -1], in quanto $6 + 2 = 8$, $6 - 2 = 4$ e $2 - 1 = 3$, $2 + 1 = 3$. Quindi la trasformazione è data da [6 2] e [2 -1].

Data una funzione $f(x)$, la sua decomposizione rispetto agli spazi V^j e W^j è definita da un'espansione sulle relative basi:

$$V^j f = \sum_{k \in \mathbf{Z}} \lambda_{j,k} \phi_k^j(x) \quad \text{con} \quad \lambda_{j,k} = \langle f(u), \phi_k^j(u) \rangle \quad (2.22)$$

$$W^j f = \sum_{k \in \mathbf{Z}} \gamma_{j,k} \psi_k^j(x) \quad \text{con} \quad \gamma_{j,k} = \langle f(u), \psi_k^j(u) \rangle \quad (2.23)$$

Dato che le funzioni di scaling e wavelet sono definite a partire dalle funzioni

madre secondo indici di risoluzione e dislocazione spaziale è possibile ottenere ricorsivamente i coefficienti di espansione $\lambda_{j,k}$ e $\gamma_{j,k}$ direttamente:

$$\lambda_{j,k} = \sum_{n \in \mathbf{Z}} g_{n-2k} \lambda_{j+1,k} \quad (2.24)$$

$$\gamma_{j,k} = \sum_{n \in \mathbf{Z}} h_{n-2k} \lambda_{j+1,k} \quad (2.25)$$

dove g h sono detti rispettivamente filtri di scala e wavelet e, nel caso di trasformate Haar, agli operatori:

$$g = \frac{1}{2}\{1, 1\} \text{ e } h = \frac{1}{2}\{1, -1\}$$

Per concludere, il procedimento sopra descritto che permette la decomposizione di un segnale viene detto **Trasformata Wavelet**, dove il tipo di trasformata (noi abbiamo preso in considerazione quelle Haar) viene definito dalla specifica dei filtri di scala e wavelet.

2.2.4 Trasformata Wavelet bidimensionale

La rappresentazione comune di un'immagine è quella di matrici di pixel, come definito all'inizio del paragrafo. È necessaria allora una estensione della teoria vista poc'anzi, per il caso bi-dimensionale.

Sono due le tecniche di applicazione delle Trasformate Wavelet : *trasformazione standard* e *trasformazione non standard*. La prima consiste nell'applicazione del processo descritto per il caso monodimensionale, alle righe fino ad arrivare al livello desiderato, ed in seguito alle colonne. Quella *non standard* procede alternativamente su righe e colonne prima di passare al livello successivo.

In figura 2.7 mostra un esempio esplicativo nel quale si evidenziano le differenze fra le due tecniche.

Nella trasformazione standard, le funzioni base, sono i possibili prodotti tensoriali delle funzioni di scala e wavelet di base del caso monodimensionale, come mostrato in figura 2.8.

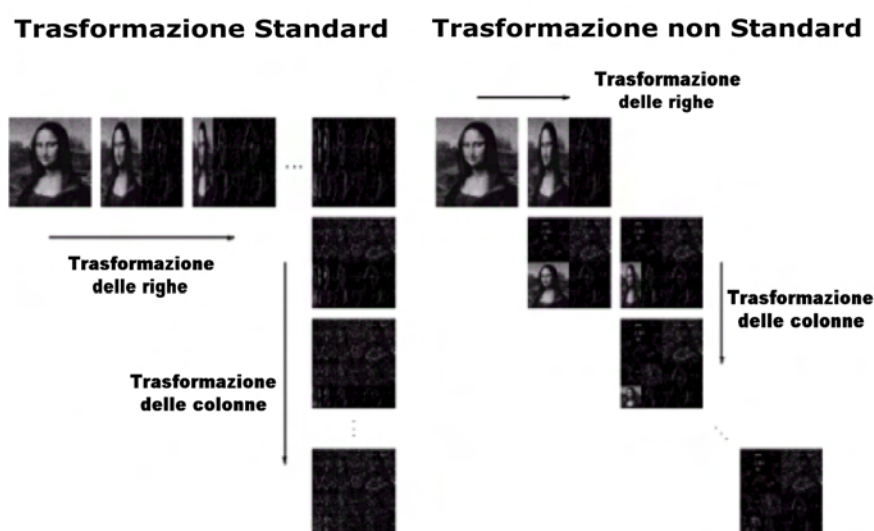


Figura 2.7: Esempio di trasformata standard e non standard

Per quella non standard è necessario ridefinire il sistema formale.

Si avrà una funzione di scala bidimensionale:

$$\phi\phi(x, y) = \phi(x)\phi(y)$$

e tre funzioni Wavelet

$$\phi\psi(x, y) = \phi(x)\psi(y) \quad \text{Coefficiente verticale}$$

che codifica differenze di intensità luminosa tra i pixel adiacenti in verticale

$$\psi\phi(x, y) = \psi(x)\phi(y) \quad \text{Coefficiente orizzontale}$$

che codifica differenze di intensità luminosa tra i pixel adiacenti in orizzontale

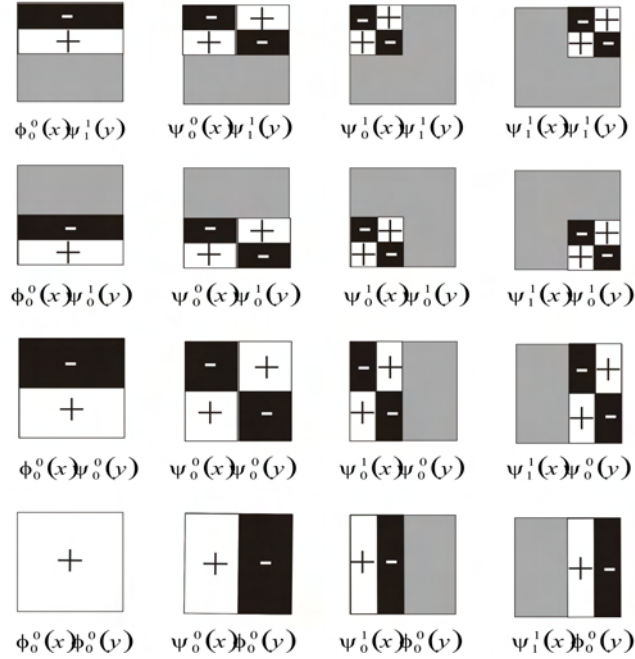


Figura 2.8: Costruzione standard di basi Wavelet Haar bidimensionali per lo spazio V^2

$$\psi\psi(x, y) = \psi(x)\psi(y) \quad \text{Coefficiente diagonale}$$

che codifica differenze di intensità luminosa tra i pixel adiacenti in diagonale

Un esempio di costruzione non standard di una base è mostrato in figura 2.9.

Negli esempi visti nelle figure 2.8 e 2.9 sono indicati anche i coefficienti per la base Haar: al segno più corrisponde +1, a quello meno -1 e alla parte vuota 0.

È difficile fare un'analisi di quale sia migliore fra le due trasformazioni, standard e non standard, presentate. C'è da considerare che la prima è molto

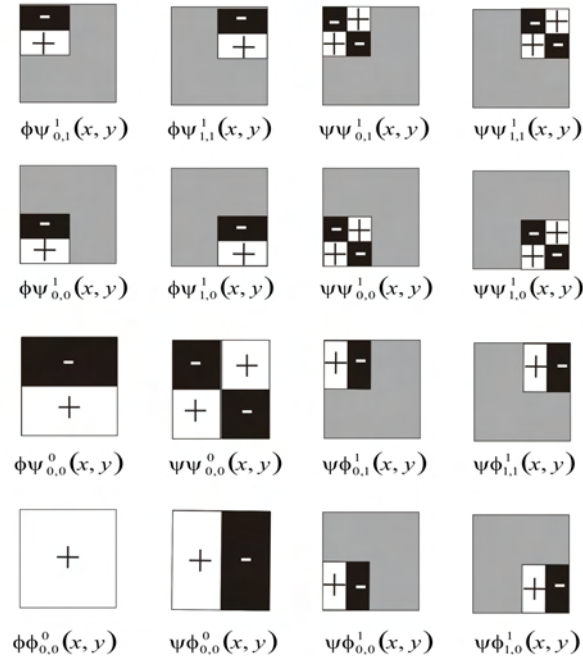


Figura 2.9: Costruzione non standard di basi Wavelet Haar bidimensionali per lo spazio V^2

semplice, basta applicare le funzioni monodimensionali, ma utilizzando la seconda ad ogni passo elabora un quarto dei coefficienti del passo precedente, contro la metà di quella standard, che di conseguenza risulta meno efficiente.

Come nella maggior parte dei problemi, la soluzione dipende dalla natura del problema stesso.

2.2.5 Trasformata Wavelet Overcomplete

Nel modello che è stato presentato fino ad ora, le funzioni di base traslano orizzontalmente e verticalmente, senza sovrapposizioni spaziali, cioè si muovono di una quantità pari al loro supporto. Per aumentare la risoluzione spaziale di

decomposizione è possibile utilizzare le cosiddette *Trasformate Wavelet Over-complete*. L'idea è di avere una codifica dei dati ridondante data da una sovrapposizione spaziale delle funzioni di base, sia di scala che Wavelet. In questo modo ad ogni porzione dell'immagine viene attribuito un numero di coefficienti più elevato.

Si definisce densità il massimo numero di funzioni che ad un dato livello di decomposizione possono essere diverse da zero nello stesso elemento del dominio. Si parla di densità singola quando la traslazione è pari al misura del supporto, doppia quando è pari alla metà della misura del supporto e quadrupla, invece, quando è un quarto (figura 2.10).



Figura 2.10: Funzioni di scala Haar a densità singola, doppia e quadrupla

Capitolo 3

Il CAD

3.1 Premesse

Il carcinoma mammario si presenta sulla mammografia come una zona generalmente molto chiara, con forma più o meno circolare e bordo più o meno definito. Nella prima parte del lavoro sono state evidenziate le innumerevoli difficoltà che ancora oggi il radiologo incontra nel diagnosticare tale disturbo, le quali sono causa di percentuali di errore altissime (10%-30%).

Si capisce che, per una malattia dove le possibilità di sopravvivenza sono strettamente legate alla sua diagnosi in fase embrionale, queste probabilità sono assolutamente troppo elevate.

Alcuni studi indicano come l'affiancare di un sistema automatico CAD, al radiologo, faccia diminuire notevolmente la possibilità di errore, senza comunque perdere di generalità.

Nonostante si siano raggiunti ottimi risultati, tanto che alcune cliniche nel mondo ne fanno già uso, la ricerca scientifica in questo campo è ancora in una fase molto sperimentale. Per cui non esistono degli standard di progettazione di tali sistemi.

Nonostante ciò se ne evidenziano alcune caratteristiche comuni principali: in

generale gli algoritmi di rivelazione di lesioni tumorali possono essere raggruppati in tre categorie differenti:

- Algoritmi composti da una fase nella quale si esegue una prima discriminazione fra le zone sospette e quelle di tessuto normale, seguita dall'estrazione di *feature* per caratterizzare le regioni trovate ed infine una fase di classificazione nella quale si apporta la selezione finale. In quest'ultima in genere ci si appoggia a sistemi di classificazione automatici ed addestrabili come l'**SVM** da noi utilizzato (**classificatori**).
- Algoritmi che estraggono le *feature* per tutte le aree dell'immagine ed utilizzando un classificatore per individuare le masse
- Algoritmi che individuano le aree sospette comparando le mammografie della mammella destra e mammella sinistra.

In questo lavoro analizzeremo solo i primi due casi.

Le *feature* rivestono un ruolo molto importante in quanto sono i dati sui quali andrà ad agire il classificatore. Devono quindi essere scelte nel modo più caratterizzanti possibile, senza però perdere di generalità.

Vista la complessità della classe di oggetti che si va ad analizzare questo compito a volte può risultare molto arduo, tanto che molti algoritmi si specializzano esclusivamente su alcuni tipi di masse.

A riguardo, una tecnica che si sta facendo largo nella ricerca scientifica è quella di appoggiarsi a sistemi di estrazione automatica delle *feature* quali, ad esempio, le **trasformate Wavelet** viste nel Capitolo 2.

3.2 Architettura del sistema

Il CAD qui presentato ha come scopo quello di individuare le eventuali masse tumorali che possono essere presenti in una mammella, partendo da una immagine mammografica digitalizzata. Una cosa importante da sottolineare è

che il sistema non deve solo avere altissime prestazioni sull'individuazione dei segnali positivi, ma altrettante nel riconoscere i negativi e quindi scartarli.

Il sistema si colloca come ibrido fra le prime due categorie di algoritmi, secondo la classificazione data nel precedente paragrafo. Considerando la difficoltà intrinseca della classe di oggetti da analizzare e considerando che spesso questi presentano caratteristiche molto simili all'ambiente che li ospita (tessuto mammario), potrebbe risultare estremamente difficile scegliere un insieme di *feature* tali da creare un modello che sia rappresentativo. Nel CAD presentato è stato implementato un algoritmo di estrazione delle caratteristiche di una massa attraverso i coefficienti wavelet.

Tali caratteristiche saranno l'input per un classificatore SVM, il quale, opportunamente addestrato, individuerà l'eventuale presenza e posizione della lesione, considerando così la rilevazione di masse come un problema di *pattern recognition a due classi*.

Si può notare come, in una visione d'insieme, non sia necessario alcun di nessun tipo di regole formali o basi di conoscenze fornite *a priori* per poter effettuare l'analisi, ma impara direttamente attraverso l'insieme di esempi dati in input.

Da un punto di vista software si è cercato di eseguire la progettazione e l'implementazione seguendo una divisione in moduli così da poter abilitare e disabilitare parti senza la necessità di metter mano nuovamente al codice. Questo risulta estremamente utile nelle fasi di sperimentazione e di aggiornamento dei moduli. L'attuale versione del sistema è composto da un insieme di librerie la cui abilitazione è possibile attraverso file di configurazione, esentando dalla necessità di ricompilare il tutto.

Uno dei grossi vantaggi di questo tipo di progettazione è la portabilità: può risultare un brutto inconveniente dover ricompilare il codice per fare prove, su macchine che non hanno tutto il corredo software necessario.

Come mostrato nella figura 3.1, il CAD è composto da cinque moduli principali:

- *Pre-detection*
- *Pre-elaborazione ed estrazione dei crop*
- *Estrazione feature (coefficienti Wavelet) per ogni crop*
- *Classificatore (SVM)*
- *Visualizzazione dei risultati*

che verranno esaminati nel dettaglio nel seguito del Capitolo.

Nelle premesse fatte in precedenza sono emerse due modalità di utilizzo del CAD, che corrispondono anche alle due fasi di elaborazione di cui è composto l'intero processo: l'addestramento (*training*) e la rilevazione (*detection*). Quella di addestramento è la fase nella quale vengono forniti alla SVM gli esempi necessari per imparare a riconoscere le lesioni tumorali. La rivelazione è invece quella nella quale dato in input un mammogramma, il sistema fornisce in output le masse tumorali che ha individuato.

Nello schema di figura 3.1 sono indicati i suoli passi logici necessari alla fase di *detection*. Per quanto riguarda la fase di addestramento, l'unica differenza sostanziale, da un punto di vista dei moduli, è che invece di utilizzare la SVM in classificazione, lo si userà per la produzione del *modello*.¹ Per una migliore spiegazione si rimanda al Paragrafo 3.5.

¹File contenente le informazioni apprese in fase di training, secondo certi parametri di configurazione, utilizzato in fase di classificazione.

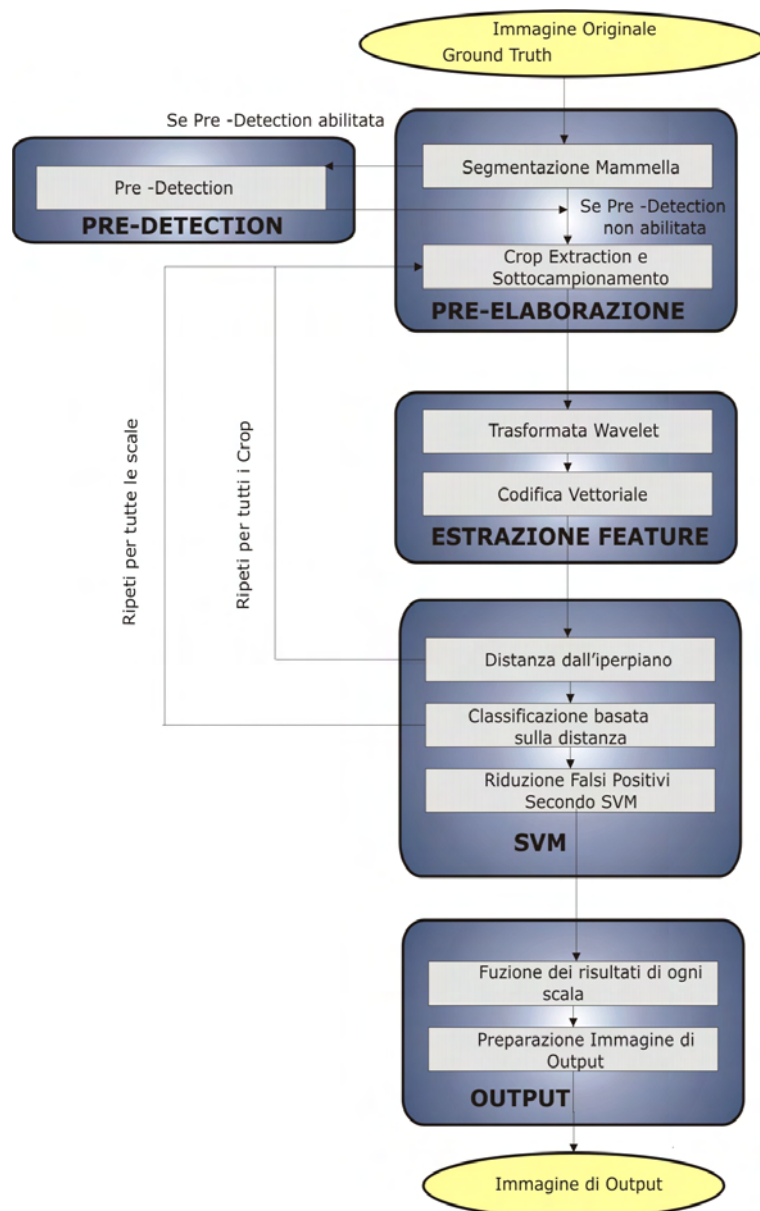


Figura 3.1: Schema dei moduli e del flusso del CAD per la fase di detection. I riquadri in grigio rappresentano le librerie che eseguono le operazioni descritte nei riquadri interni

3.3 Pre-elaborazione

La pre-elaborazione è composta da tre fasi principali:

- *Segmentazione*
- *Ridimensionamento dell'immagine segmentata*
- *Estrazione dei crop*

3.3.1 Segmentazione

Il CAD prende in input il mammogramma digitalizzato (che chiameremo immagine originale) ed una immagine relativa al Ground Truth, che è l'indicazione spaziale della presenza di masse. Se ne vedrà meglio in seguito l'utilità.

Come si può osservare dagli esempi di mammografie presentati nel primo capitolo (figure 1.2-1.3-1.4-1.5) non tutta l'immagine è occupata dalla mammella, ma vi sono parti, che al nostro scopo sono assolutamente inutili e che devono quindi essere eliminate. Le motivazioni sono le seguenti:

- La parte complementare alla mammella comprendere informazioni che in un qualche modo potrebbero confondere il classificatore: presentando oggetti, come ad esempio l'etichetta, che si appartengono alla classe dei negativi, ma molto distanti dall'iperpiano di separazione, si condiziona la buona definizione della classe stessa.
- Le immagini mammografiche sono scannerizzate ad altissima risoluzione, si parla di dimensioni che vanno da 4000x2000 a 5000x3000 pixel, per cui ogni minima elaborazione è computazionalmente non trascurabile. Soprattutto vista l'importanza che la velocità di diagnosi riveste come parametro di valutazione del sistema. Quindi, eseguire operazioni su parti non interessanti è un dispendio di tempo assolutamente inutile.

La segmentazione si occupa di tagliare queste zone superflue, generando un file contenente le coordinate della zona nella quale risiede la mammella (figura 3.3) e sulle quali verranno poi eseguite le elaborazioni successive. La 3.3 rappresenta un codifica in formato grafico del suddetto file di segmentazione.

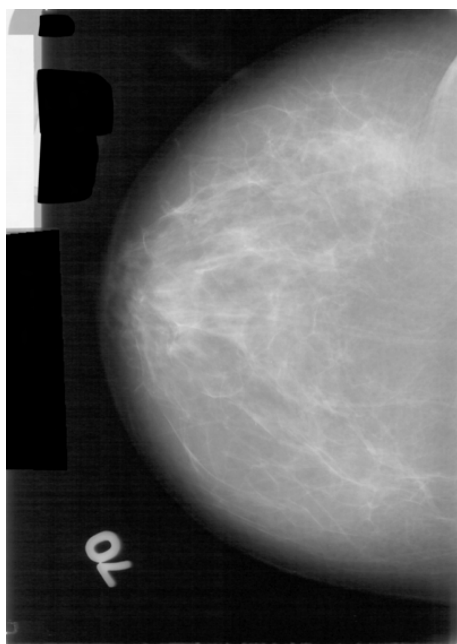


Figura 3.2: Immagine originale

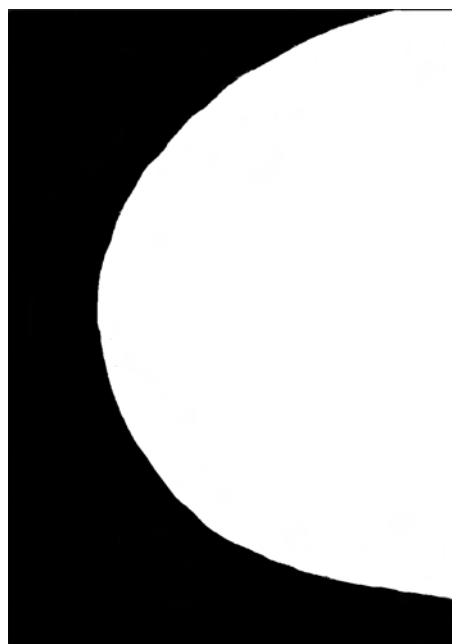


Figura 3.3: Immagine segmentata

3.3.2 Ridimensionamento ed Estrazione dei Crop

Il modulo prende in input l'immagine mammografica e le informazioni relative alla segmentazione, prodotte al passo precedente.

Il suo compito è quello di generare un insieme di riquadri dell'immagine originale, sulle quali poi eseguire l'estrazione delle feature. A tale scopo una maschera quadrata a dimensioni variabili viene fatta passare su tutta la zona segmentata con un passo di scansione fissato al 10% della dimensione lineare della maschera stessa, come mostrato in figura 3.4. L'effetto voluto è quello

di avere sovrapposizioni fra i riquadri. Scelta dettata da vincoli di omogeneità del set di esempi positivi che vengono presentati alla SVM. Come si vedrà in seguito, un segnale positivo è stato scelto come una porzione dell'immagine contenente completamente una massa, la quale è centrata nel riquadro. Il classificatore è quindi addestrato a riconoscerla in questa forma. Senza la sovraesposizione molte lesioni si perderebbero in quanto non inquadrare nel giusto modo.

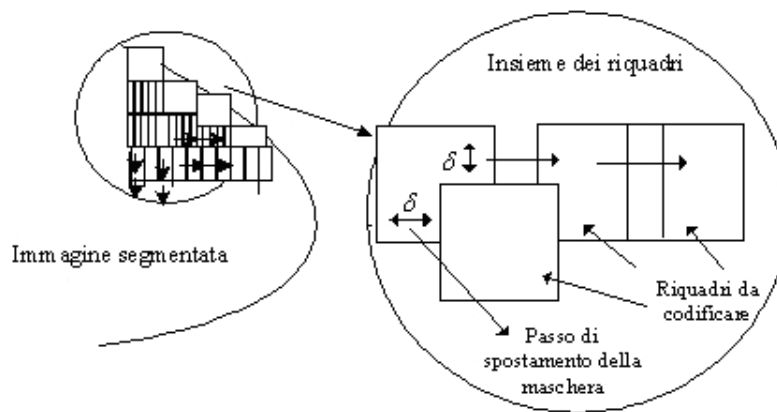


Figura 3.4: Movimento della maschera di scansione

A causa del vincolo di omogeneità sorge un ulteriore problema: le lesioni hanno una dimensione che varia fra 3mm e 30mm. La dimensione del vettore delle caratteristiche, che viene dato in input al classificatore, è estremamente dipendente dalla dimensione del riquadro di scansione. Il vincolo dimensionale posto dal classificatore si traduce anche in vincolo fisico sull'estensione dei riquadri prodotti: tutti i riquadri scansionati devono essere della stessa grandezza. Il problema che si delinea, allora, è come estrarre le zone con una classe di oggetti da rilevare così dimensionalmente eterogenea.

La soluzione implementata è la seguente. Vengono eseguite diverse fasi di scan su tutta l'immagine segmentata, variando le dimensione della maschera estraendo serie di riquadri fra loro dimensionalmente differenti. Dopo di che

vengono tutti ridimensionati a 64x64 pixel. Il processo viene eseguito attraverso un algoritmo di subsampling ad interpolazione bilineare.

La scelta di fissare la maschera di riferimento a 64x64 pixel è stata fatta sulla base del miglior compromesso fra perdita di informazioni ed efficienza computazionale.

Esempio.

Si consideri una immagine di input di 4000x3000 pixel con grandezza di un pixel di $50\mu m$ e siano $32mm$ (640 pixel), $16mm$ (320 pixel) e $10mm$ (213 pixel) le dimensioni delle masse cercate. Vengono eseguiti degli scan con maschere di riquadri di 640x640 pixel, 320x320 pixel ed infine 213x213 pixel. Ridimensionando tutti i crop estratti rispettivamente del 10%, 20% e 30% si ottiene l'effetto desiderato di avere tutte le zone di 64x64 pixel.

L'immagine segmentata, come si vede in figura 3.3, per le caratteristiche geometriche della mammella, non ha bordi rettangolari, mentre la maschera di scansione sì. Viene alla luce allora un ulteriore problema: la sola scansione della parte segmentata porta al rischio di perdita di lesioni tumorali o per mancato centramento o addirittura per mancato rientro in un crop. Si vedano figure 3.5 e 3.6.

Il problema è stato risolto con un piccolo accorgimento. Nel momento in cui la maschera arriva sul bordo, continua l'avanzamento fino a che il proprio centro non ha lo stesso valore di coordinata longitudinale del punto di intersezione fra la base del riquadro e la segmentazione della mammella. La figura 3.7 illustra il procedimento.

Il modulo restituisce in output i crop ridimensionati necessari per la codifica al passo successivo. Di questi, come si vedrà in seguito, è di fondamentale importanza non perdere le coordinate spaziali e le informazioni di scala, i quali serviranno in fase di ricomposizione per la visualizzazione dell'immagine finale diagnosticata. Essi vengono quindi salvati in particolari strutture dati.

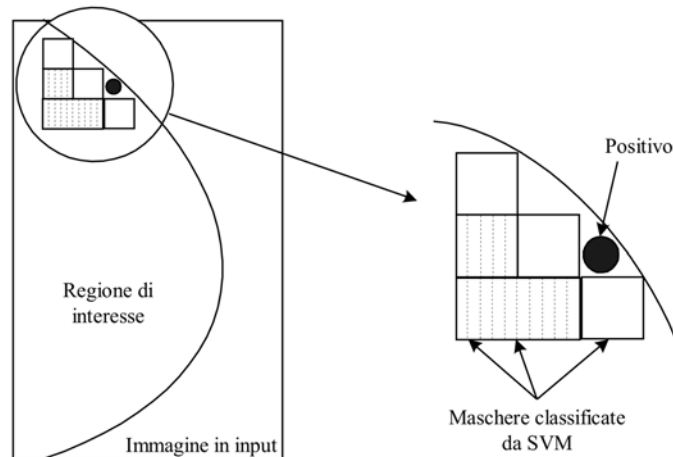


Figura 3.5: Esempio di lesione persa dovuto alla scansione interna alla segmentazione

3.4 Estrazione Feature

Il seguente modulo prende in input i crop estratti nella fase precedente codificati come matrici di punti luminosi.

È composto da due parti:

- *Trasformata Wavelet*
- *Codifica vettoriale*

3.4.1 Trasformata Wavelet

Come visto nel Capitolo 2 le trasformate Wavelet descrivono l'immagine espressa secondo matrice di valori di luminosità, in una serie di coefficienti, detti appunto *coefficienti wavelet*, che ne danno una rappresentazione nella quale sono evidenziate la struttura e la morfologia. Tali coefficienti sono di tre tipi, ognuno dei quali codifica le differenze di luminosità dei pixel di una zona lungo le

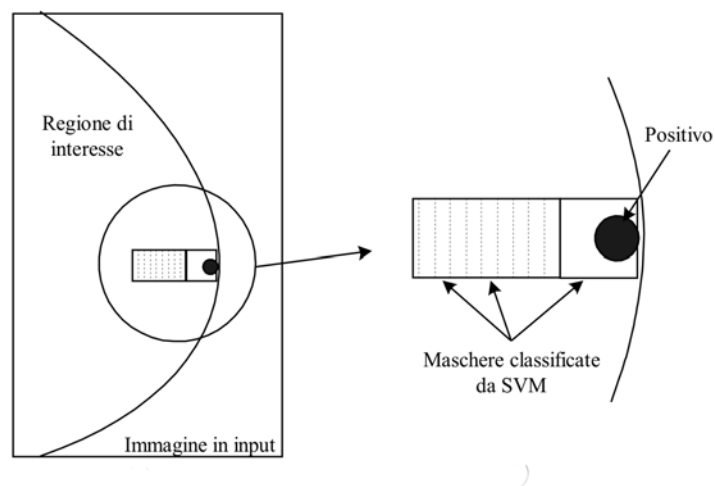


Figura 3.6: Esempio di lesione rilevata male dovuto alla scansione interna alla segmentazione

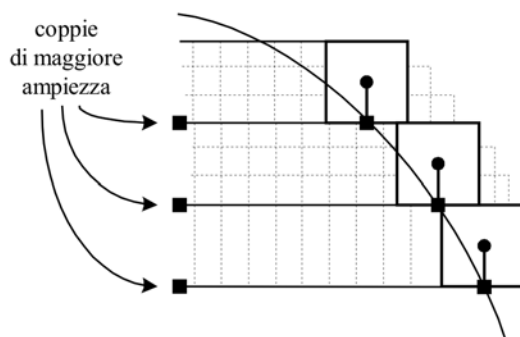


Figura 3.7: Estrazione dei crop sul bordo della segmentazione

tre direzioni: orizzontale, verticale ed obliquo. Per cui risulteranno coefficienti alti laddove le differenze sono alte, bassi dove il tessuto è uniforme. Situazioni di questo si hanno in genere quando in quel punto l'immagine presenta un bordo di un oggetto.

Vista la complessità della classe di oggetti da classificare, risulta molto difficile individuare grandezze misurabili che li possano caratterizzare senza limitare la possibilità del sistema di generalizzare: la scelta di feature da parte dell'uomo sarebbe comunque vincolata e ristretta alla sua visione delle masse tumorali. Per questo ci si è affidati ai coefficienti delle Trasformate Wavelet.

Per ogni crop estratto nella fase di pre-elaborazione vengono calcolati i Coefficienti Wavelet a diversi livelli di decomposizione. Da alcuni dati empirici, ottenuti a diverse fasi di sperimentazione e, supportati da nozioni di base teorica, è emerso che le Trasformate Wavelet più idonee al problema da risolvere sono le Haar a densità doppia.

Come già detto in precedenza i crop sono in formato matriciale bidimensionale; questo comporta che ad ogni livello dello sviluppo wavelet si hanno tre matrici: una dei coefficienti orizzontali, un dei verticali ed una di quelli diagonali. Di conseguenza, preservando tre livelli, alla fine dell'elaborazione si avranno nove matrici di coefficienti.

Di seguito viene riportato uno schema che sintetizza la sequenza di decomposizione per crop di 64x64 pixel di dimensione:

- **Livello 1** → estrazione di 63(righe)*63(colonne)*3(direzioni) coefficienti
- **Livello 2** → estrazione di $62 * 62 * 3$ coefficienti
- **Sottocampionamento** → sottocampionamento dei livelli di scala mantenendo solo i coefficienti di ordine dispari. Ne rimangono 31.
- **Livello 3** → estrazione di $30 * 30 * 3$ coefficienti
- **Livello 4** → estrazione di $29 * 29 * 3$ coefficienti
- **Sottocampionamento** → sottocampionamento dei livelli di scala mantenendo solo i coefficienti di ordine dispari. Ne rimangono 14.
- **Livello 5** → estrazione di $13 * 13 * 3$ coefficienti
- **Livello 6** → estrazione di $12 * 12 * 3$ coefficienti

Di tutti i livelli si mantengono solo i coefficienti livello 2,4 e 6. ne deriva che in totale, per ogni crop, si hanno: 14487 coefficienti.

Il valore trovato deve poi essere moltiplicato per il numero di crop, che pure non è assolutamente trascurabile. Ne consegue allora una gestione, da parte del sistema, di una elevatissima quantità di dati e da ciò si comprende l'importanza dei tempi di elaborazione. La scelta di adottare le trasformate Haar è proprio condizionata dalla loro semplicità e conseguente velocità di calcolo.

La fase di sottocampionamento apportata ogni due livelli ha il duplice scopo di abbassare il numero di coefficienti, che altrimenti sarebbe ingestibile, e di mantenere un effetto di multirisoluzione. Dimezzando, infatti, il numero di coefficienti di scala, al livello successivo due coefficienti wavelet successivi raccolgono informazioni da coefficienti di scala relativi ad una zona più ampia dell'immagine.

Il modulo prende in input una serie di crop generati nella fase di pre-elaborazione. Questi possono essere passati direttamente in un unico passo, oppure salvati ed acquisiti in file dove ognuno rappresenta una zona. L'aggiunta di questa possibilità si rivelerà molto utile in fase di apprendimento della SVM (Paragrafo 3.5)

3.4.2 Codifica Vettoriale

Nella fase di estrazione delle feature vengono prodotte nove matrici per ogni riquadro. Dato che la forma nella vengono passati dati alla SVM è quella vettoriale, in particolare un vettore per ogni crop, si richiede una fase di trasformazione delle matrici in vettori compatibili.

Tali vettori sono costruiti nel seguente modo:

- Ogni matrice viene copiata, facendo una scansione per riga, in un vettore.
- I vettori di cui al punto 1, vengono uniti secondo il seguente ordine: raggruppati per direzione e ordinati per livello

Il vettore risultante che verrà poi dato al classificatore ha la seguente forma:

$$\left(\underbrace{\text{Livello2, Livello4, Livello6}}_{\text{Orizzontale}} - \underbrace{\text{Livello2, Livello4, Livello6}}_{\text{Vetricale}} - \underbrace{\text{Livello2, Livello4, Livello6}}_{\text{Diagonale}} \right)$$

3.5 Addestramento

La fase di addestramento ha lo scopo di fornire al classificatore una base di conoscenza, secondo la quale essere in grado di risolvere un problema di pattern matching. Questo viene fatto presentando una serie di esempi, significativi, delle due classi considerate: nel nostro caso positivi e negativi.

Si è scelto di usare SVM, come classificatore per la sua capacità a generalizzare con relativamente pochi esempi di training. Sono sufficienti infatti solo alcune migliaia di immagini per avere come risultato un buon addestramento. Risultano notevoli i vantaggi a cui porta ciò, sia da un punto di complessità computazionale, con vantaggi sui tempi di calcolo, che di complessità logica del sistema: un sistema più semplice è più facilmente controllabile, e manutenibile.

Sulla base del concetto di ottimizzazione fatta sulla massimizzazione della distanza fra gli iperpiani di confine fra le due classi (Vedi 2.1.2), la SVM utilizza solamente i Vettori di Supporto per la generalizzazione del problema. Questo condiziona pesantemente la scelta dell'insieme di esempi di training. È necessario fornire immagini che possano ben definire i cosiddetti *bordi delle classi*.

È noto come il Pattern Recognition in mammografia sia problematico a causa dell'alto grado di eterogeneità degli oggetti in esame. Per cui si deve scegliere un set di esempi che tengano anche conto di ciò; deve cioè essere rappresentativo di ogni tipo di lesione. Per questo motivo, nello specifico del problema di Pattern Recognition per lesioni tumorali al seno, vengono costruiti set di esempi con un numero di negativi di gran lunga superiore a quello dei positivi, in quanto variabilità molto più elevata.

Il modulo implementato prende in ingresso una serie di vettori generati nella fase di estrazione delle feature, quali esempio delle due classi, e prova a risolvere il problema dell'ottimizzazione dell'iperpiano.

Siccome, in fase di pre-elaborazione, il trattamento delle immagini è leggermente differente se queste servono per l'addestramento, rispetto alla normale detection; è necessario fare un passo indietro per analizzare l'interno processo. Allo scopo di discriminare fra segnali positivi e negativi da passare alla SVM, nella fase iniziale, oltre al mammogramma viene anche caricata un'immagine bitmap in due scale di grigio, detta *Ground Truth*, tutta nera esclusa le zone con masse tumorali. Queste sono derivate da mammogrammi nei quali il radiologo ha precedentemente segnato le lesioni tumorali (vedi Figure 3.8 e 3.9).

È durante la fase di estrazione dei crop che avviene la discriminazione fra gli esempi positivi e quelli negativi. Date le informazioni aggiuntive fornite dal Ground Truth è immediata la localizzazione delle lesioni. Per cui la loro estrazione avviene ritagliando un riquadro più grande della dimensione lunga della massa, del 30% e centrato su di essa. Chiaramente per vincolo di omogeneità imposta dalla SVM viene poi ridimensionato a 64x64 pixel.

I negativi sono tutto ciò che non è positivo. Per cui viene fatta una scansione della parte segmentata della mammografia, con maschere di dimensione fissa, o con la tecnica di multiscala, e tutti i riquadri estratti vanno a far parte dell'insieme dei negativi. In particolare, per evitare che questi contengano parti di masse, si è adottati la tecnica di estrazione dei positivi da mammogrammi con lesioni, mentre i negativi da quelli raffiguranti mammelle sane, quindi senza masse tumorali.

Dei crop ne viene poi fatta la codifica Wavelet, senza nessuna differenza rispetto al caso già analizzato, e passati separatamente, positivi e negativi, al primo SVM per l'addestramento.

Allo scopo di raffinamento del modello, si introduce una tecnica di *bootstrap* automatico. L'idea è la seguente. Da un bacino di immagini, ne viene scelto un set usato per il training del primo SVM e un set detto di validazione

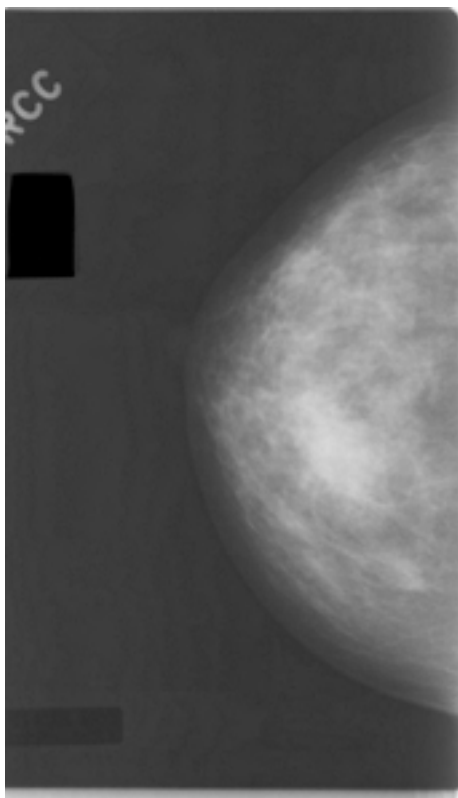


Figura 3.8: Immagine Originale

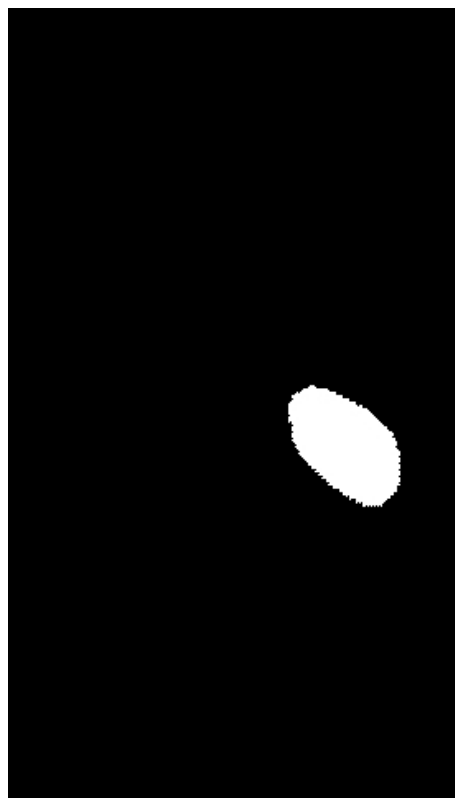


Figura 3.9: Ground Truth

(*validation*) che viene utilizzato per la verifica della bontà dell'addestramento. Una volta creato un modello, questo viene testato eseguendo una detection su una parte del set di Validation, quindi delle quali il classificatore non ha alcuna informazione. Questo viene fatto allo scopo di osservare la risposta reale: eseguendo test su immagini delle quali la SVM ha già acquisito informazioni, ci si aspetta che le classifichi quasi perfettamente. Risulterebbe quindi alquanto inattendibile. Da questo test verranno prodotti dei segnali falsi positivi, i quali, completeranno il set iniziale di training per un nuovo addestramento. Questo processo può essere ripetuto diverse volte, fino a che i risultati non sembrano soddisfacenti. Per ulteriori chiarimenti si veda la figura 3.10

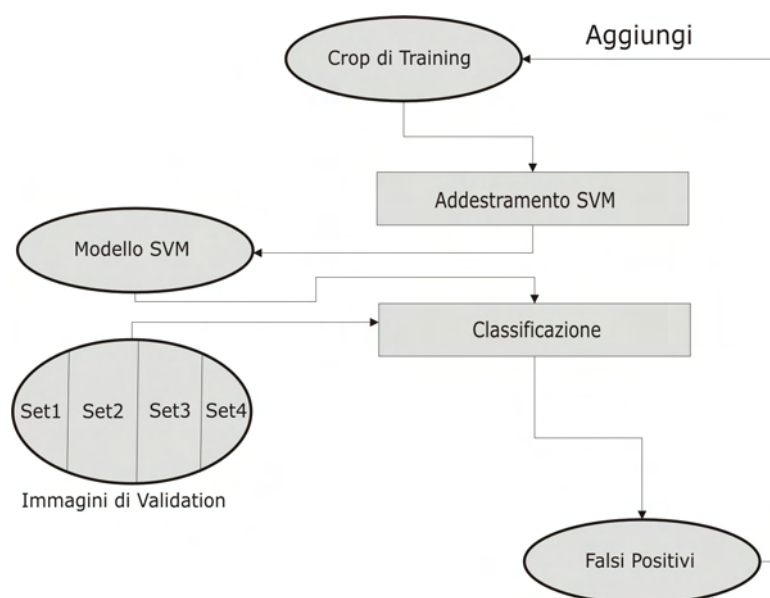


Figura 3.10: Fase di Addestramento del primo SVM

Il secondo classificatore viene addestrato con i veri positivi del set di training della prima e completato con quelli del set di validation. Per quanto riguarda gli esempi negativi da usare, si considerano i falsi positivi dell'ultimo test di classificazione eseguito dalla prima SVM. L'idea è di segnare i veri positivi che riconosce anche il primo, ma rimediare ai suoi errori.

3.6 Riconoscimento

La fase di riconoscimento consta di due classificazioni poste in serie, eseguite da altrettanti SVM addestrati in modo differente. Per le modalità di addestramento di entrambi si rimanda al paragrafo 3.5.

Allo scopo di classificare un oggetto, rappresentato da un vettore di coeffi-

cienti, un SVM ha necessità di un base di conoscenza, la quale gli viene data dal modello generato in fase di apprendimento. Per cui il primo passo è prelevare le informazioni da esso, sotto forma di triplette: (x_i, α_i, y_i) , dove x_i sono i vettori di supporto usati nel modello, α_i i relativi moltiplicatori di Lagrange e y_i la loro classe di appartenenza.

Una volta fatto ciò la SVM è pronto ad essere usato per una detection, per cui, dato un nuovo vettore, si procederà alla sua classificazione secondo la relazione 2.15, la quale restituirà un valore 1 o -1 in base alla risultante classe di appartenenza decisa.

Il passo successivo è quello di calcolare la distanza dell'elemento considerato dall'iperpiano, con lo scopo di assegnargli un indice di forza, detto *indice di confidenza*, definito come il grado di correttezza della classificazione. L'idea è che un segnale classificato positivo ha più probabilità che lo sia realmente, tanto più esso è distante dall'iperpiano di separazione. Per cui tanto più alto è il valore dell'indice di confidenza.

In una prima fase di sperimentazione il sistema utilizzava un solo classificatore SVM. Si è visto in seguito che ponendone un secondo in cascata a quello già esistente, ed opportunamente addestrato, poteva ridurre il numero di falsi positivi.

Quindi l'attuale implementazione comprende un primo SVM che riceve in input la serie dei vettori, rappresentanti i crop composti nella fase di estrazione delle feature, e restituisce in output quelli classificati come positivi. Da sottolineare che i segnali considerati negativi dalla SVM da ora in poi non vengono più presi in esame, in quanto non sono di utilità: tutto ciò che non è positivo, è negativo.

I candidati sopravvissuti alla prima fase di classificazione vengono passati al secondo SVM, il cui compito è di eliminare i *falsi positivi* rimasti. Si è notato che in un qualche modo vi è una certa ridondanza fra i segnali che vengono classificati erroneamente: è probabile che nella maggior parte dei casi il primo SVM sbaglia sempre sullo stesso tipo di segnali.

Per cui diventa logico addestrare il secondo a riconoscere questi, in modo da

eliminarli (vedi paragrafo 3.5).

Per dare una quantificazione dell'insieme di dati, in genere al primo classificatore arrivano circa 1000 crop ($1000 * 14487$ coefficienti), dei quali ne rimangono 50 che vengono processati dal secondo, per concludere con 4-5 segnali.

3.7 Visualizzazione dell'Output

I vettori di coefficienti classificati positivi dal secondo SVM corrispondono a zone del mammogramma. Per cui vengono individuati i crop corrispondenti, non scalati a 64×64 , ma nella forma originale. A causa della scansione in multiscala, vi è la possibilità che diversi riquadri si riferiscano alla stessa zona, ma chiaramente con i centri spostati. Viene allora fatta una unione, portandoli tutti alla dimensione originale e facendo un'operazione OR: l'idea è di mantenere tutte le masse individuate.

3.7.1 Unione di classificatori

Come dimostrato da alcune ricerche([17]), classificatori assolutamente indipendenti, addestrati a risolvere lo stesso tipo di problema di Pattern Recognition hanno la tendenza ad individuare gli stessi *Veri Positivi* e differenti *Falsi Positivi*. Sulla scia dei risultati di tali studi è nata l'idea di applicare la tecnica al sistema CAD implementato, con l'intento di, ancora una volta, abbassare il numero di falsi positivi, utilizzando diverse configurazioni dello stesso, il più possibile indipendenti. La variazione, che ha portato alla generazione di modelli differenti, è avvenuta sul set e sui parametri di addestramento.

Attualmente il sistema prevede la possibilità di combinare tre differenti classificatori.

I risultati di ognuno sono uniti secondo la politica *votazione due su tre*: un segnale è mantenuto se almeno due dei tre CAD individuano la zona, considerando un certo margine di sovrapposizione (figura 3.11).

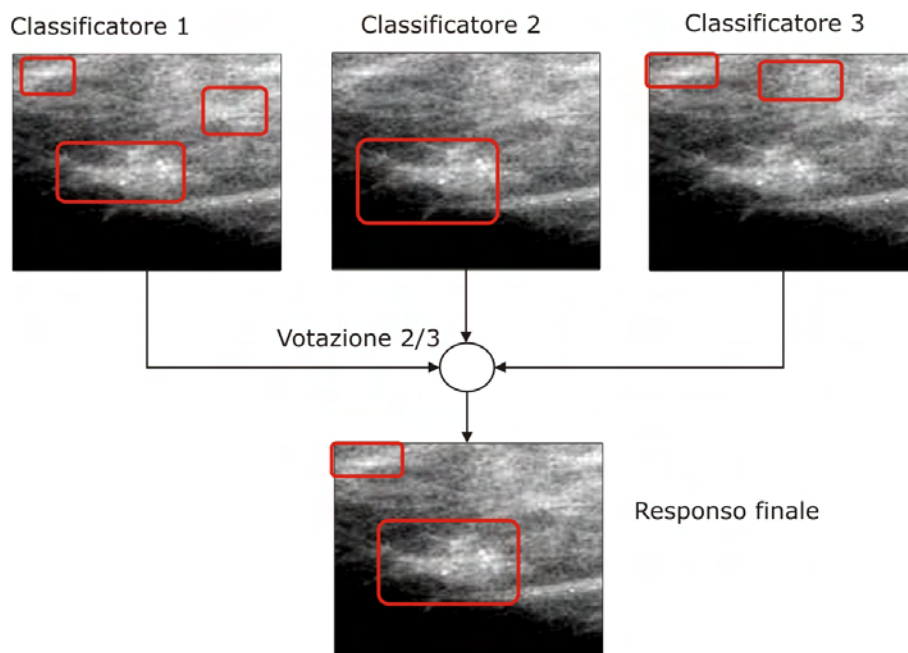


Figura 3.11: Combinazione dei risultati di diversi sistemi esperti

3.8 Il CAD parallelo

Una immagine mammografica ha una dimensione lineare che varia da 4000x2000 a 5000x3000 pixel. Considerato il numero elevatissimo di operazioni che devono essere eseguite su di essa, ci si accorge che i tempi di elaborazione possono essere molto alti. Soprattutto se si considera l'importanza che la velocità di risposta riveste in questi tipo di sistemi: la lentezza implica abbandono dell'uso. Su di una macchina moderna seriale monoprocesso, il CAD, con l'architettura e il metodo di funzionamento indicati fino ad ora, necessita di un tempo di elaborazione che ammonta a circa 3 minuti per immagine. L'analisi completa di ogni paziente comprende quattro mammogrammi, da cui si deducono i tempi complessivi: circa 12 minuti.

Le richieste attuali in ambito ospedaliero indicano come accettabile un tempo di risposta del CAD di un minuto per immagine.

Si comprende che diventano fondamentali metodologie di ottimizzazione!

L'idea è che vi sono molte parti dell'elaborazione che possono essere parallelizzate in quanto lavorano su dati non condivisi. Un esempio è il seguente. Non è necessario che l'estrazione dei coefficienti Wavelet dei riquadri avvenga in modo sequenziale. Esse sono informazioni assolutamente indipendenti. Si potrebbe pensare a diverse macchine in parallelo che elaborano ognuna un differente crop.

Sono stati utilizzati diversi livelli di parallelismo, in dipendenza delle macchine che andranno ad ospitare il software:

- SSE ²
- Thread su SMP ³
- MPI ⁴

Il massimo grado di parallelismo sfruttabile con il CAD implementato si ottiene utilizzando un cluster di computer, fra loro interconnessi, ognuno dei quali è biprocessore. Il cluster ha una struttura di tipo Beowulf ⁵, nella quale è prevista l'esistenza di un Master. A questo vengono affidati i seguenti compiti:

- esecuzione di tutte le parti che devono essere eseguite in seriale in quanto non parallelizzabili
- la gestione dei servizi del cluster

²Intel Streaming SIMD Extension. Set di operazioni incluse dalla Intel dal Pentium III in poi

³Symmetric Multi Processing. Piattaforme multiprocessore forniti di un numero di processori variabile da 2 a 64 i quali condividono la memoria

⁴Message Passing Interface. Standard per il paradigma Message Passing

⁵Architettura a multicomputer che può essere usata per calcoli paralleli. È un sistema che normalmente consiste di un nodo server e uno o più nodi client connessi via Ethernet o altri tipi di rete.

- la distribuzione dei task e dei dati ai vari nodi di calcolo (nel sistema analizzato consistono in elaboratori del cluster)
- raccogliere e comporre i risultati delle elaborazioni di ogni singolo nodo per fornire l'output finale

Considerando che il Master una volta eseguita la divisione dei compiti, rimarrebbe inattivo fino al momento della raccolta dei dati, esso stesso diventa un nodo.

Creando la suddetta divisione del lavoro si potranno evitare inutili colli di bottiglia che possono rallentare il calcolo.

Le comunicazioni fra il Master e le altre macchine del cluster avviene attraverso il paradigma MPI.

L'intera struttura è ben evidenziata in figura 3.12.

Nella fase iniziale il Master si occupa dello scanning del mammogramma con le Maschere e divide i crop risultanti, fra tutti i nodi slave del cluster, preoccupandosi di dar loro anche il modello SVM utilizzato in classificazione.

All'interno di una stessa macchina, nel caso venga supportato da un punto di vista Hardware, avviene una divisione in Thread. Per ora il sistema prevede l'utilizzo di macchine al massimo biprocessore. Quindi sono creati due thread diversi ad ognuno dei quali è assegnato un diverso crop da elaborare. Per definizione dei Sistemi Operativi, i thread hanno memoria condivisa; questo li rende molto veloci in fase di inizializzazione. Inoltre sono altamente ottimizzati grazie all'utilizzo di SMP.

Chiaramente la divisione non verrà fatta se la macchina è monoprocessore.

È anche possibile eseguire tutto il sistema su una unico SMP, quindi senza considerare clustering, eliminando la parte di MPI.

L'ottimizzazione della SVM viene effettuata utilizzando le operazioni di base

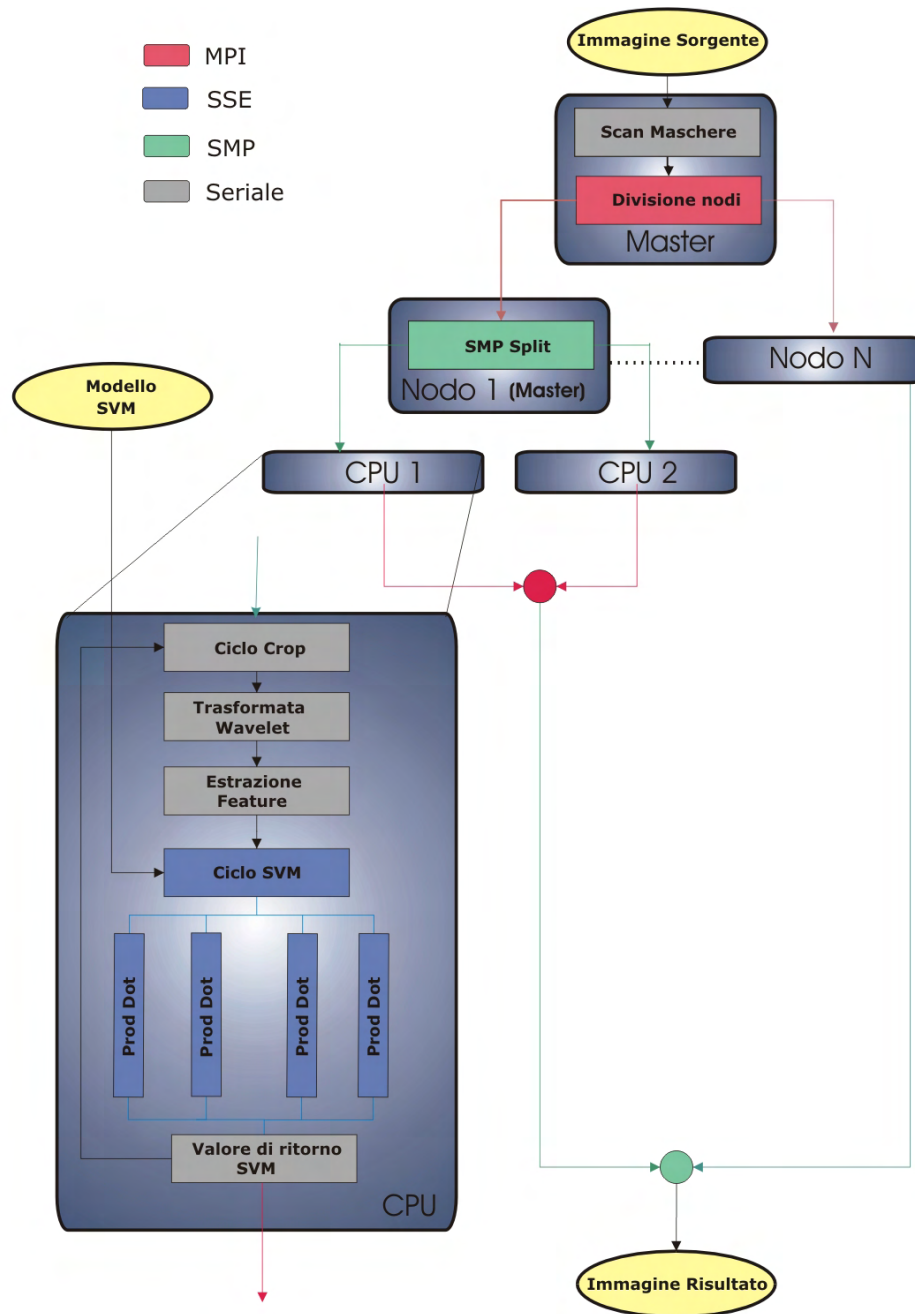


Figura 3.12: Struttura del CAD con tutti i livelli di ottimizzazione. La label Master nel Nodo 1 sta ad indicare il fatto che una volta eseguita la divisione, egli stesso diviene un nodo.

SSE. Se si considera la 2.15 si nota come il problema della classificazione non è altro che un prodotto matriciale fra la matrice del modello (Vettori di Supporto) e quella delle feature (tutti i vettori di ogni singolo crop vengono raggruppati in una matrice).

SSE implementa già una operazione di moltiplicazione vettoriale, il cui uso nel sistema rende molto più veloce un calcolo estremamente oneroso.

Per concludere adottando queste ottimizzazioni su Pentium III a 1000 Mhz si raggiunge un tempo di elaborazione per immagine da 120 a 12 secondi in base al numero di nodi adottato.

Capitolo 4

Pre-detection

4.1 Premesse

Pur essendo di natura molto varia, il tessuto mammario si può dividere in due tipologie fondamentali:

- zone grasse, che sono radio-lucenti
- zone ghiandolari o fibrose, come i vasi sanguigni, radio-opache

Queste ultime, in particolare, si contraddistinguono in quanto a grado di luminosità sulla lastra mammografica che risulta decisamente più elevato. Un mammogramma, quindi, da un punto di vista morfologico, si presenta con un fondo abbastanza uniforme su toni scuri e con una parte ben strutturata, più in evidenza, su toni decisamente più elevati. Grazie alla composizione tessutale generalmente densa del carcinoma mammario, la loro risposta al fascio di raggi X è simile a quella dei corpi radio-opachi, con risultante alta luminosità della zona sul mammogramma.

Da quanto detto si deduce che le difficoltà di localizzazione delle lesioni tumorali non sussistono nei casi in cui il tessuto ospitante che le circonda è di composizione grassa. I veri problemi nascono nel momento in cui, invece, nascono in un tessuto estremamente denso. È in queste situazioni, infatti, che

per caratteristiche simili le lesioni possono essere confuse per tessuto sano o viceversa.

Nel precedente capitolo si è descritta la fase di segmentazione, nella quale dal mammogramma vengono asportate le zone esterne alla mammella, le quali non sono di interesse al nostro scopo, in quanto non possono essere lì localizzate masse tumorali. Basandosi sull'analisi della distribuzione del colore appena fatta, unita al concetto stesso di segmentazione, ci si pone la seguente domanda: esistono zone anche all'interno della parte rappresentante la mammella, nelle quali vi è una bassissima possibilità di trovare lesioni?

Dalla caratteristica del carcinoma di essere collocato su frequenze spettrali molto alte, nasce l'idea che sta alla base del modulo di pre-detection: segmentare la parte interna alla mammella, eliminando quelle zone, di tessuto, aventi frequenze spettrali basse, nelle quali vi è una alta probabilità di non trovare masse.

L'output del modulo di pre-detection è dato da una immagine binaria con indicate le Regioni di Interesse (ROI), nelle quali è possibile l'esistenza di una massa (vedi figura 4.4). Grazie a questa, la successiva fase di scanning delle maschere non verrà più eseguita su tutta l'area del seno, ma solo su queste zone.

I benefici portati dall'utilizzo di tale tecnica sono principalmente due:

- diminuzione del tempo di calcolo dovuto alla minore area di applicazione dei processi successivi
- migliore definizione della classe dei negativi. La SVM risolve lo MMH problem considerando solamente i Support Vector, che sono gli elementi ai margini delle classi e, di conseguenza, quelli più simili fra loro. Si ha che la fase di pre-detection tende ad eliminare i negativi scontati, mantenendo solamente quelli simili a masse, cioè quelli con maggiori

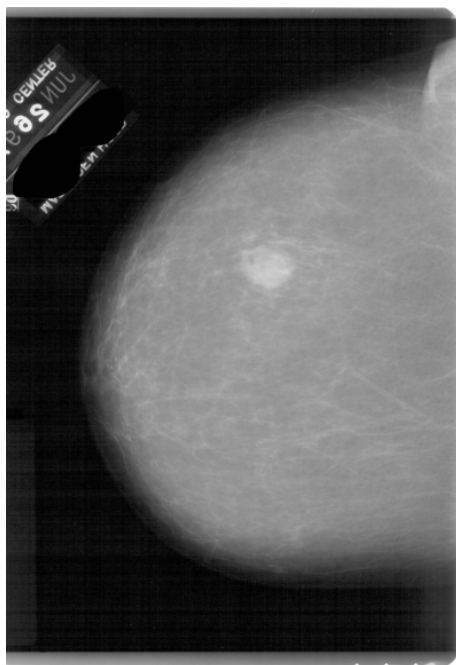


Figura 4.1: Immagine input

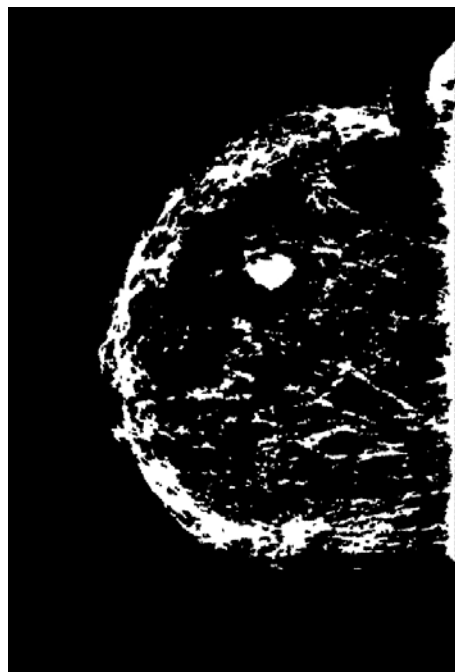


Figura 4.2: Immagine output

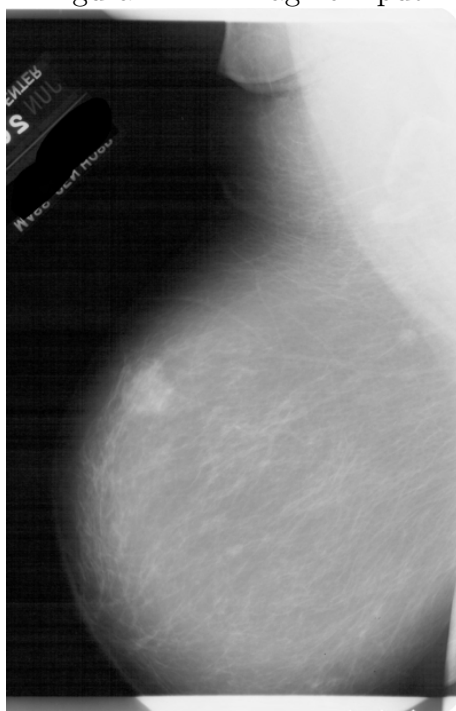


Figura 4.3: Immagine input

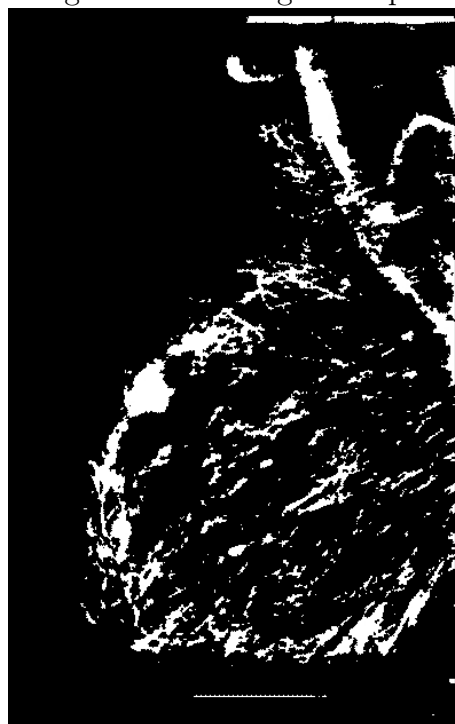


Figura 4.4: Immagine output

probabilità di diventare Support Vector. La conseguenza è una migliore definizione della classe dei negativi.

Nel capitolo successivo si vedrà come queste supposizioni siano supportate da una fase di sperimentazione che ne dimostra la validità.

Il modulo di *pre-detection*, come è visibile dallo schema completo del CAD in figura 2.1, è posto come parte opzionale nelle fasi preliminari di tutto il processo. È opzionale in quanto può essere abilitato o disabilitato senza problemi di sorta.

L'intero processo di pre-detection consta delle seguenti fasi:

- *Riduzione di scala*
- *Filtro Passa Alto*
- *Sogliaatura (Threshold)*
- *Applicazione Operatori Morfologici*
- *Aumento scala alle dimensioni originali*

I processi logici sono descritti nel grafico rappresentato in figura 4.5.

4.2 Ridimensionamento

Allo scopo di non perdere dettagli utili, i mammogrammi vengono acquisiti utilizzando scanner ad altissima definizione, generando immagini con risoluzioni spaziali che variano fra i 4000x2000 e i 5000x3000 pixel. Come si vedrà in seguito, la fase di pre-detection è costituita dal susseguirsi di elaborazioni che necessitano la scansione dell'intera immagine, dalle cui dimensioni, chiaramente, è in diretta dipendenza il loro tempo di esecuzione.

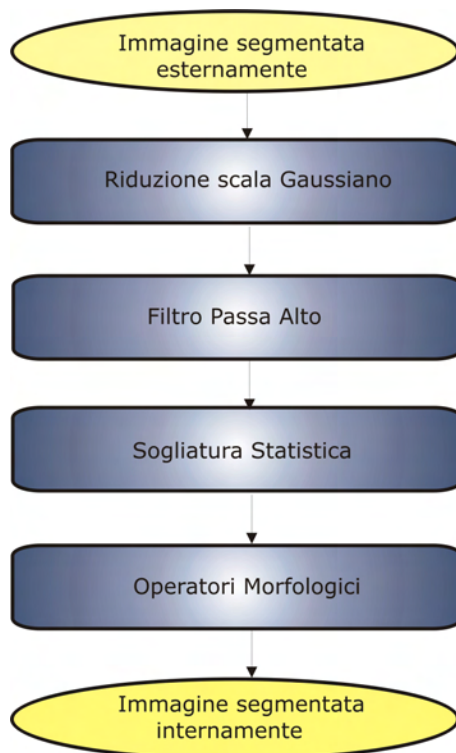


Figura 4.5: Schema del flusso della fase di pre-detection

Si può intuire che l'applicazione di tali processi sulle scale originali diventerebbe assolutamente inaccettabile. Per cui viene introdotto come primo gradino di pre-detection, una riduzione delle dimensioni del mammogramma. Ma, come è intuibile, ad ogni processo di ridimensionamento a valori inferiori ne consegue una diminuzione delle capacità informative dell'oggetto dimensionato, a causa della eliminazione di parte dei suoi pixel.

Non sempre, però, vi possono essere solo conseguenze negative. Come già detto in precedenza, infatti, ogni processo di digitalizzazione introduce segnali di disturbo; ne risulta che una perdita d'informazione sull'immagine, implica anche un abbassamento del grado di rumore. Anzi, come si vedrà in seguito saranno proprio questi i pixel più colpiti.

Viene messo in luce, allora, il secondo aspetto di questa fase: l'eliminazione dei segnali indesiderati presenti nel mammogramma. Il processo verrà spiegato in modo più accurato nei prossimi paragrafi.

4.2.1 Convoluzione

Le tecniche di filtraggio applicate all'elaborazione delle immagini (*image processing*), si suddividono in due grandi famiglie: filtri nel dominio spaziale e filtri nello dominio delle frequenze. I primi basati sulla propria applicazione direttamente ai valori di luminosità dei pixel, i secondi agiscono sulle frequenze spettrali.

Come si può intuire queste ultime possono risultare decisamente costose, in quanto necessitano di un passaggio dal normale dominio spaziale di codifica dell'immagine a quello delle frequenze e viceversa, attraverso particolari trasformate come quella di Fourier (*Fast Fourier Transform*).

Vista l'importanza dei tempi di elaborazione in un sistema CAD, si è scelti di adottare algoritmi appartenenti alla prima classe citata, cioè quelli nel dominio spaziale.

Si rende necessaria ora una più rigida formalizzazione del problema.

È sempre possibile descrivere un'immagine come una funzione bidimensionale che permette la mappatura dello spazio delle coordinate di un pixel con il proprio valore di luminosità:

$$f(x, y) = k \quad \text{con } x \text{ e } y \text{ coordinate dell'immagine}$$

Si definisce *trasformazione* di una immagine $f(x, y)$, l'immagine $g(x, y)$ risultato dell'applicazione di una funzione T ad $f(x, y)$:

$$g(x, y) = T(f(x, y)) \quad \text{con } x \text{ e } y \text{ coordinate dell'immagine}$$

Alla base degli algoritmi di filtraggio si ha il concetto di *convoluzione*, che in questa sede viene trattato solo nel dominio spaziale. Data una matrice, detta *matrice di convoluzione*

t_1	t_2	t_3
t_4	t_5	t_6
t_7	t_8	t_9

e data una immagine $f(x, y)$, si definisce *convoluzione* la trasformazione:

$$T(f(x, y)) = t_1 * f(x-1, y-1) + t_2 * f(x, y-1) + t_3 * f(x+1, y-1) + t_4 * f(x-1, y) + t_5 * f(x, y) + t_6 * f(x+1, y) + t_7 * f(x-1, y+1) + t_8 * f(x, y+1) + t_9 * f(x+1, y+1)$$

Applicata ad ogni pixel dell'immagine.

È da notare che nel caso in cui i valori della matrice siano maggiori dell'unità esiste una grossa possibilità che nel processo di convoluzione si abbia un errore di overflow sul valore massimo dei toni di grigio. Per evitare ciò essa viene normalizzata a 1. Cioè la somma dei suoi valori deve risultare 1.

I concetti esposti possono essere estesi anche a matrici di convoluzione di dimensioni differenti (da 1x1 alla grandezza dell'immagine), comunque dispari, per ovvie ragioni.

In conclusione, la base di ogni processo di filtraggio è la stessa. Ciò che differisce è la definizione e la grandezza della matrice di convoluzione. La struttura di un generico algoritmo è la seguente:

```

for (i=1 to Num_row_image)
  for (j to Num_col_image)
  {
    image[i][j]=0
    for (k=(-Dim_mat/2) to (Dim_mat/2))
      for (m=(-Dim_mat/2) to (Dim_mat/2))
      {
        Ni = i+k
        Nj = j+m
        if Ni,Nj inside image
          image[i][j] = image[i][j]+mat[Ni][Nj]
      }
  }
}

```

4.2.2 Filtro Passa Basso Gaussiano

Alla base dell'intero processo di ridimensionamento vi è un *Filtro Passa Basso Gaussiano*. Come tutti i filtri nel dominio spaziale, esso si realizza attraverso la convoluzione con una matrice, dalla quale, per la sua particolare composizione a *campana di Gauss*, prende il nome l'algoritmo stesso.

Tale matrice viene riportata di seguito.

$$\frac{1}{4} * \begin{array}{|c|c|c|} \hline \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \hline \frac{1}{2} & 1 & \frac{1}{2} \\ \hline \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \hline \end{array}$$

dove $\frac{1}{4}$ è la costante di normalizzazione.

Le metodologie di applicazione sono le seguenti. Si consideri per ora una riduzione di scala che dimezza le dimensioni lineari dell'immagine. Quindi la nuova matrice di codifica avrà un'area di un quarto rispetto alla originale (Figura 4.6).

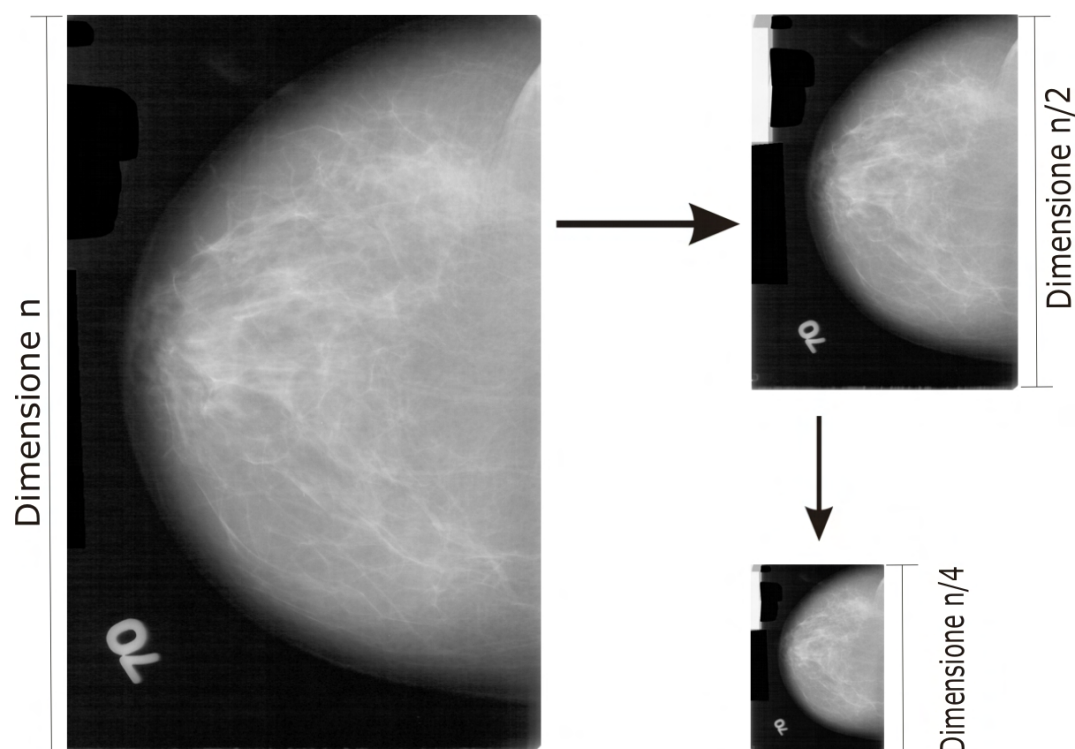


Figura 4.6: Riduzione di una immagine, prima di $\frac{1}{4}$, poi di $\frac{1}{8}$ alla dimensione originale.

Il valore di luminosità di un pixel della nuova immagine deriva dal processo di convoluzione dell'originale con la matrice sopra definita. Allo scopo di dimezzare linearmente lo spazio delle righe e quello delle colonne, il processo non viene applicato ad ogni pixel, ma ad uno ogni due (Vedi Figura 4.7).

Questo comporta che due valori di luminosità di pixel vicini saranno in parte determinati dalla convoluzione di valori comuni. Lo scopo di tale sovrapposizione risulta quello di addolcire gli spostamenti della matrice, creando una relazione fra valori contigui nella immagine prodotta.

Si può notare come sia possibile anche un dimensionamento su scale maggiori di quelle di 2: si potrebbe, ad esempio, fare una riduzione di scala con dimen-

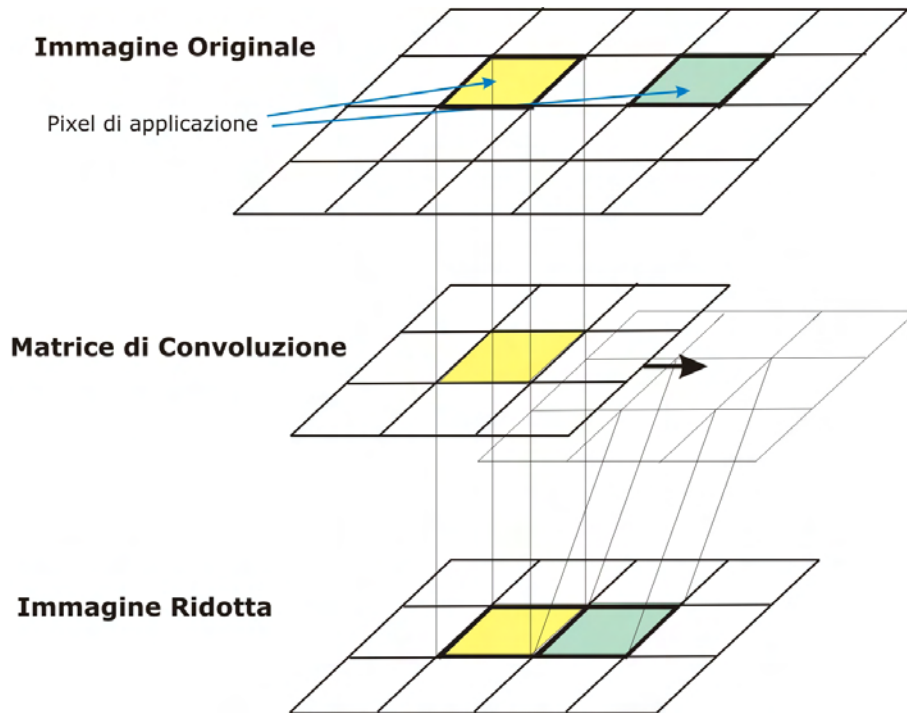


Figura 4.7: Convoluzione per la riduzione applicando il filtro passa basso gaussiano.

sioni lineari $\frac{1}{3}$ di quelle originali. Per fare ciò è sufficiente aumentare a 5×5 le grandezza della matrice di convoluzione, con un passo di spostamento del centro di 2 pixel.

Nel sistema CAD presentato è data la possibilità di eseguire n volte la riduzione Gaussiana, applicando lo stesso algoritmo in cascata, in modo che il mammogramma prodotto da uno sia l'input per l'altro. Si ottiene così facendo un ridimensionamento lineare di 2^n dove n è un parametro definibile dall'utente. I valori tipici sono 2 o 3.

Il processo convolutivo, sotto certi accorgimenti, non è altro che una media pesata di valori di luminosità. In particolare, è una media secondo i pesi di una

distribuzione Gaussiana. La conseguenza del processo è un livellamento dei valori, con conseguente eliminazione dei cosiddetti picchi, cioè di quelle zone con luminosità molto più levata rispetto al circondario. Per questo motivo viene chiamato filtro passa basso o *smoothing filter*.

In particolare quest'ultimo nome gli è attribuito dall'effetto sfumatura (*smoothness*) risultante dopo la sua applicazione. Effetto che, sotto certi aspetti può essere un fattore assolutamente negativo, in quanto porta ad un degrado della definizione dei contorni, a causa dalla sua tendenza all'appiattimento dei contrasti.

L'uso di un matrice Gaussiana non tratta tutti i pixel allo stesso modo, ma attribuisce un peso gradualmente maggiore avvicinandosi al centro. Questo ha il vantaggio di appiattire il rumore, che in genere è composto da 2-4 pixel, ed invece preservare maggiormente i contorni degli oggetti, caratterizzati da dimensioni notevolmente maggiori.

4.3 Filtro Passa Alto

Come indicato dal nome stesso, il Filtro Passa Alto, applicato all'elaborazione delle immagini, ha lo scopo di porre in evidenza le zone con frequenze spettrali alte, a scapito di tutte le altre che subiscono una conseguente attenuazione.

Il mammogramma per sua natura ha una divisione naturale fra due tipologie di tessuti che in essa coesistono, quello grasso e quello denso, localizzati, da un punto di vista spettrale, in bande di frequenze poco sovrapposte. Come già ampiamente detto, le masse tumorali sono identificate da valori di luminosità in genere elevati. Per questo motivo il passo successivo dell'elaborazione è l'applicazione di un filtro passa alto, appunto, con lo scopo di far risaltare quelle zone nelle quali è più probabilmente localizzata una lesione.

Un ulteriore beneficio che ne deriva, è la tendenza ad aumentare la definizione dei bordi di un oggetto che nel caso del sistema CAD ha un duplice effetto:

- sopperire allo smoothing introdotto nella precedente fase di Filtraggio (Filtro Passa Basso)
- dare una migliore definizione delle classi di oggetti coinvolte: positivi e negativi

L'algoritmo utilizzato nel sistema presentato è *High Boost filter*, il cui funzionamento è indicato nello schema di figura 4.8.



Figura 4.8: Schema logico dell' High Boost Filter.

L'idea generale che sta alla base dell'algoritmo segue una politica sottrattiva: per evidenziare certe particolari frequenze è sufficiente attenuare o togliere

le altre. Il procedimento, allora, è quello di applicare una prima fase di filtraggio allo scopo di ottenere le basse frequenze, chiaramente attraverso un filtro passa basso e, poi sottrarre queste all'originale.

Per sottrazione fra immagini si intende la sottrazione matematica, apportata ai valori di luminosità dei pixel corrispondenti (con le stesse coordinate) delle due immagini distinte.

La matrice di convoluzione utilizzata per il filtro Passa Basso è la seguente:

$$\frac{1}{25} * \begin{array}{|c|c|c|c|c|} \hline 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 \\ \hline \end{array}$$

In questo caso non è stata utilizzata una matrice di convoluzione Gaussiana, ma un cosiddetta *piatta* (*plain*), in quanto non è assolutamente di interesse conservare informazioni riguardo gli oggetti rappresentati, anzi è si dimostra necessario un drastico appiattimento dell'immagine. A tale scopo, prima dell'applicazione dell'algoritmo, essa viene ulteriormente scalata, attraverso il suddetto Filtro Passa Basso Gaussiano, di un fattore 2^m , dove m è un parametro definito dall'utente. Dalla sperimentazione effettuata, come sarà dimostrato nel capitolo seguente, i valori che danno i risultati migliori sono 4,5,6.

È da notare come, secondo tali parametri, l'immagine diventi veramente molto piccola. Si consideri ad esempio un mammogramma di dimensioni 5000x3000 pixel. Ponendo i valori di $n = 2$ e $m = 5$, in seguito alle riduzioni di scala, rimane un'immagine di 40x23 pixel. È evidente che la sua capacità informativa è molto bassa rispetto all'originale.

Prima di eseguire l'operazione di sottrazione fra i due mammogrammi, chiaramente, la matrice risultante dal filtro passa basso viene riscalata a dimensione

originale, attraverso un algoritmo di ridimensionamento. L'effetto desiderato è mostrato in figura 4.13.

Il mammogramma originale e quello finale del processo di filtraggi High Boost, a confronto, sono indicati rispettivamente nelle figure 4.9 e 4.10.

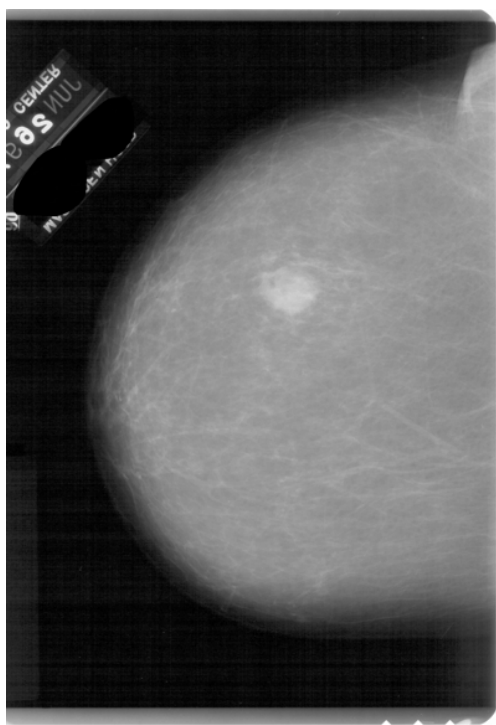


Figura 4.9: Mammogramma Originale.

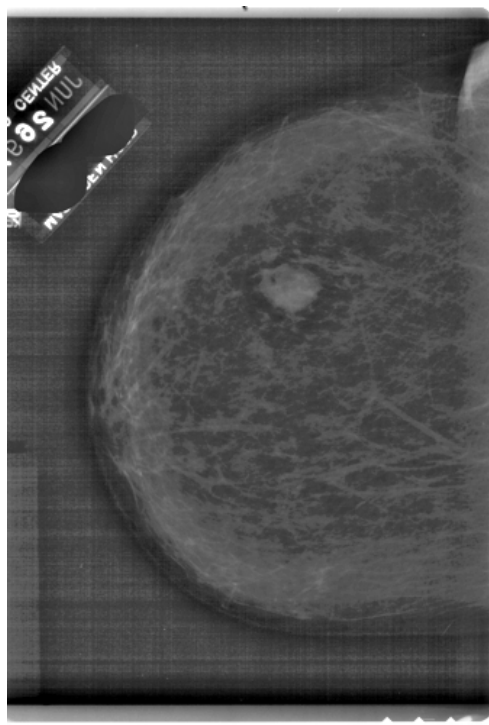


Figura 4.10: Immagine risultante di un High Boost Filter.

Dal mammogramma finale, risultato del processo di filtraggio, si può facilmente notare come si sia creata una netta distinzione fra i due differenti tipi di tessuto della mammella, aumentando il contrasto e definendo i bordi. Il fenomeno è ancora più evidente se si prende in considerazione il suo istogramma 4.12. Sono infatti molto pronunciati i due picchi, che tra l'altro non hanno elementi di intersezione: sono assolutamente distinti.

Come si vedrà questo risulterà estremamente utile nella successiva fase, quella di Sogliatura.

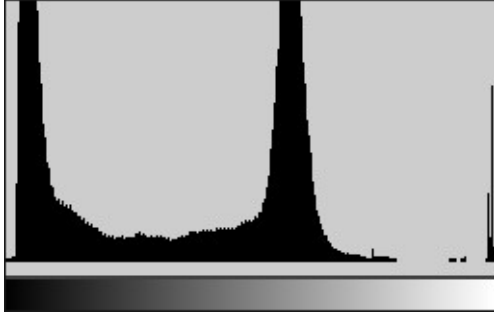


Figura 4.11: Istogramma spettrale dell'immagine originale.

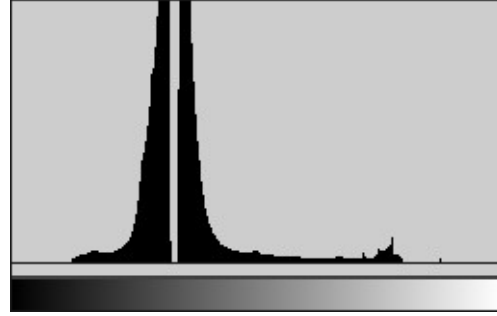


Figura 4.12: Istogramma spettrale dell'immagine di figura 4.10.

4.4 Sogliatura (Threshold)

La sogliatura dell'immagine è il cuore della fase di pre-detection. Si ricorda che lo scopo prefissato è quello di generare una immagine binaria di segmentazione interna alla mammella, nelle quale le zone mantenute siano solo quelle di interesse (ROI), dove cioè vi è una buona possibilità di localizzazione di un carcinoma.

La costruzione di tale immagine è proprio il compito della Sogliatura.

Il procedimento consiste nel prendere in input l'immagine filtrata dall'High Boost e renderla binaria applicando un taglio sull'istogramma.

Formalmente il processo è il seguente.

Data una immagine $f(x, y)$ ed il suo istogramma spettrale, si definisca un valore T , detto *soglia*, come una particolare frequenza nel range di appartenenza di $f(x, y)$. L'immagine binaria risultante è definita dalla seguente equazione:

$$bin(x, y) = \begin{cases} 1 & \text{se } f(x, y) > T \\ 0 & \text{se } f(x, y) < T \end{cases}$$

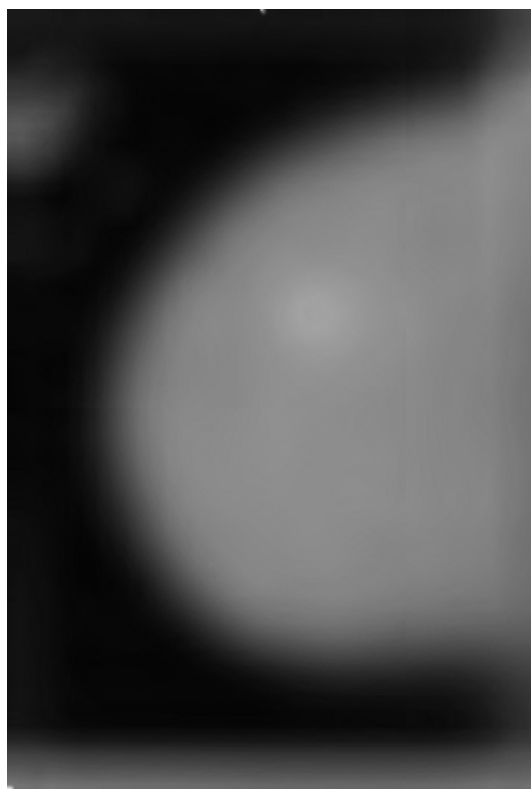


Figura 4.13: Immagine risultato dall'applicazione di un Filtro Passa Basso con matrice piatta 5x5

Il punto critico dell'intero processo di sogliatura è la determinazione del valore di soglia ottimale. I parametri di valutazione su T sono:

- *area totale delle Regioni di Interesse*: Deve essere minimizzata in quanto minore è l'area, minore è la zona di applicazione dell'estrazione dei crop nella successiva fase del CAD, minore è il tempo di esecuzione.
- *probabilità che una massa tumorale venga persa nella segmentazione*: Deve essere minimizzata in quanto una perdita in questa fase del processo, non è recuperabile nella successiva fase di classificazione.

Si osservi il mammogramma di figure 4.14 e 4.15: è abbastanza intuibile che all'aumentare di T diminuirà l'area totale delle ROI e aumenterà la possibilità di perdita delle masse tumorali.

Come verrà analizzato nel capitolo successivo, tutta la fase di sperimentazione del modulo di pre-detection, si è concentrata sulla ricerca di valori ottimali per la risoluzione del *trade off* fra l'area totale e la perdita di segnali positivi.

4.4.1 Algoritmo di Threshold

L'algoritmo di Threshold adottato nel presente sistema CAD, utilizza un valore di soglia T dinamico, calcolato su *media* (M) e *deviazione standard* (Σ) locali di un'area del mammogramma. L'intera immagine viene processata traslando una maschera di dimensioni $R \times R$, con R parametro definibile dall'utente, e calcolando ad ogni passo i valori di M e Σ .

Per evitare dislivelli troppo elevati della soglia di zone contigue, che porterebbe a differenze troppo evidenti di segmentazione, le traslazioni delle maschere avvengono con circa il 60% di sovrapposizione.

Il valore di soglia T viene determinato secondo la seguente equazione:

$$T = M + \alpha * \sigma + M * \kappa \quad (4.1)$$

dove α e κ sono definiti dall'utente e danno la possibilità di intervento sulla sogliatura. Il primo è un fattore moltiplicativo per la deviazione standard, utilizzato, in genere, per attenuare il suo effetto. Sicuramente più interessante è il parametro κ . Esso è stato introdotto in una fase di ottimizzazione dell'algoritmo per risolvere particolari reazioni indesiderate che il sistema aveva in alcune parti circoscritte del seno.

Osservando la segmentazione di figura ?? si nota che le zone ai margini della mammella, quella esterna e quella vicino al muscolo mammario, sono molto

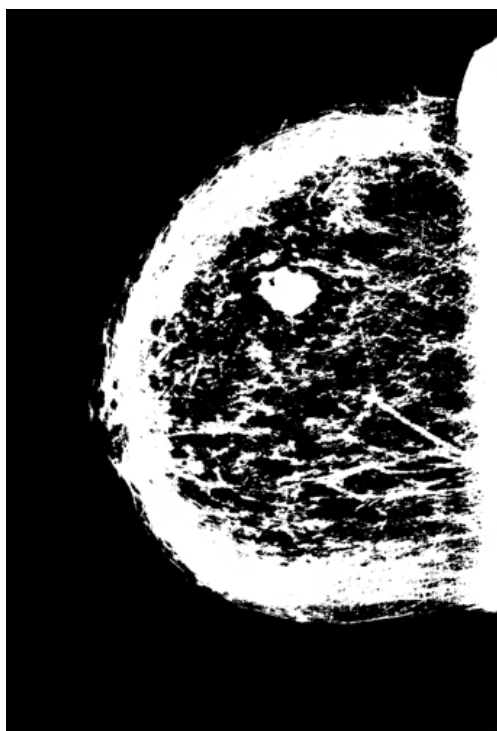


Figura 4.14: Sogliatura del mammo-gramma in figura 4.10 secondo un valore di soglia $T1 < T$, dato T il valore di soglia ottimale.

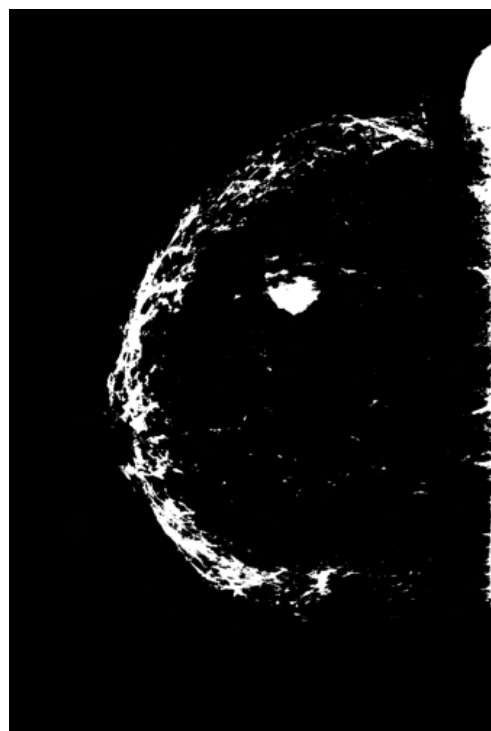


Figura 4.15: Sogliatura del mammo-gramma in figura 4.10 secondo un valore di soglia $T2 > T$, dato T il valore di soglia ottimale.

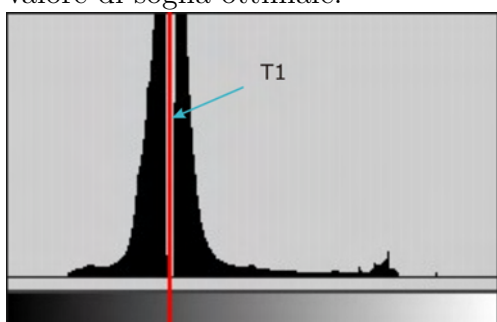


Figura 4.16: Istogramma della sogliatura $T1$.

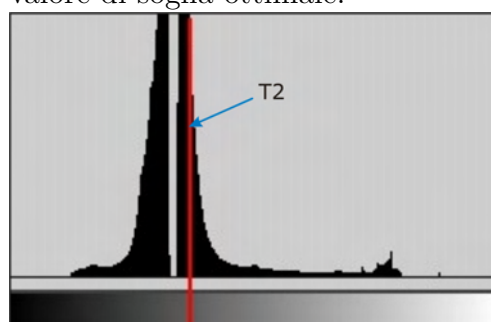


Figura 4.17: Istogramma della sogliatura $T1$.

più dense rispetto al tessuto di quella centrale. Per meglio spiegare il fenomeno si trattano separatamente i due casi, in quanto generati da problematiche differenti.

Si è notato che il tessuto vicino all'attaccatura con il busto presenta in genere una alta concentrazione di tessuto grasso, quindi con una distribuzione di luminosità decisamente uniforme. Prendendo in considerazione l'equazione 4.1, si nota che se la deviazione standard è piccola, conseguenza dell'uniformità del tessuto, senza il parametro κ , T risulterebbe molto vicino alla media. Questo porterebbe all'effetto sopra descritto, in quanto quella zone avrebbe un valore di soglia molto più basso rispetto al resto del mammogramma. Il parametro aggiunto dà la possibilità di normalizzare a piacimento T .

Esiste un altro metodo di utilizzo della sogliatura appena descritta, allo scopo di tagliare imperfezioni di digitalizzazione. Si è notato che in molti mammogrammi è presente una banda bianca, compatta, chiaramente imperfezione del processo di acquisizione. Si può pensare di applicare un threshold molto drastico, localizzato solo essa, in modo da eliminarla. Questa è la modalità di utilizzo che ne viene fatto nel presente sistema CAD.

Il problema è esattamente l'opposto per quanto riguarda il bordo esterno. In questa zona infatti, lo spessore del seno si riduce molto velocemente, per cui traspare maggiormente il tessuto ghiandolare. Questo porta chiaramente ad avere molti pixel oltre il valore di soglia. L'effetto del parametro κ è il medesimo.

La prima versione del sistema si basava su di un threshold classico, senza il parametro aggiunto e applicato in modo identico su tutta la mammella. L'ottimizzazione introdotta divide il seno nelle tre zone sopra citate, come mostrate in figura 4.18 dando la possibilità all'utente di specificare un parametro κ differente per ognuna.

La larghezza delle bande di riferimento possono essere definite dall'utente, in modo indipendente.

Chiaramente torna alla luce il *trade off* descritto prima: più area si taglia

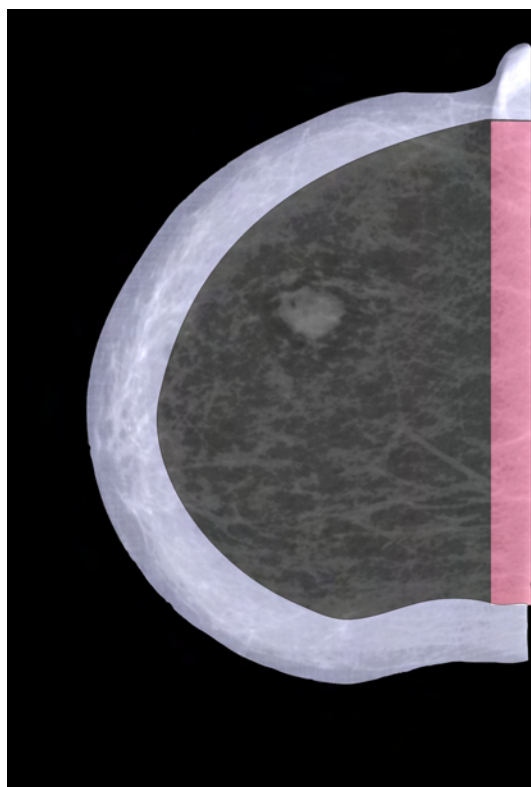


Figura 4.18: Divisione delle zone per il Threshold

più vi è possibilità di perdere masse.

In figure 4.19 4.20 e 4.21 4.22 sono riportati esempi dei suddetti tagli.

4.5 Operatori Morfologici

Gli algoritmi applicati nei vari passi della pre-detection, sono tutte elaborazioni orientate al pixel. Questo significa che non hanno coscienza della possibilità che pixel uniti formino altri oggetti di interesse. Per cui, come si può vedere dalla figura 4.22, l'immagine prodotta fino ad ora è molto sporca di pixel com-

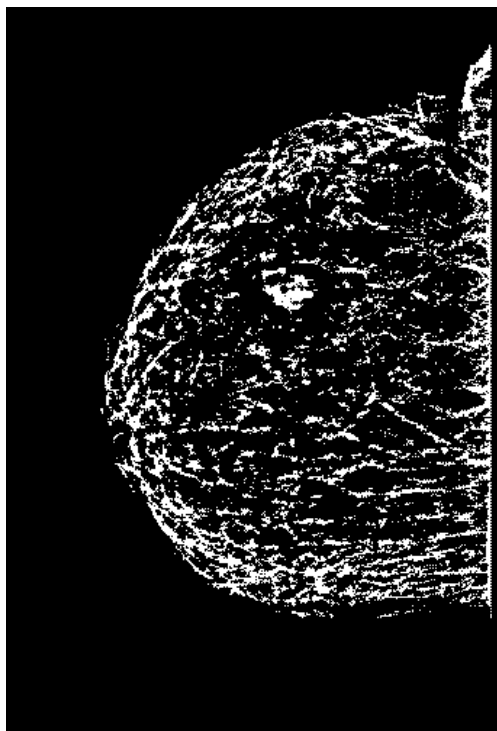


Figura 4.19: Sogliatura Omogenea su tutta la mammella.

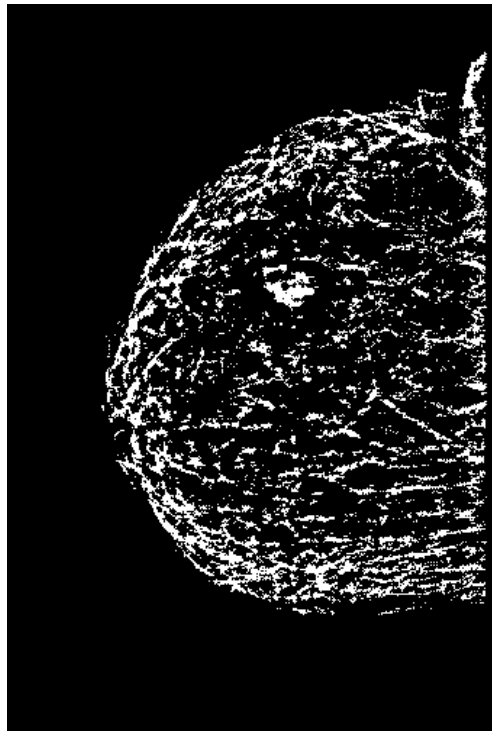


Figura 4.20: Differenziazione della sogliatura sul bordo destro.

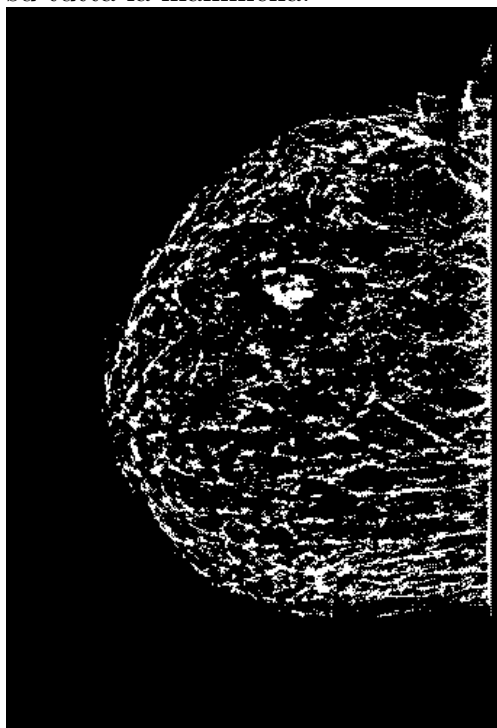


Figura 4.21: Differenziazione della sogliatura sul bordo sinistro.

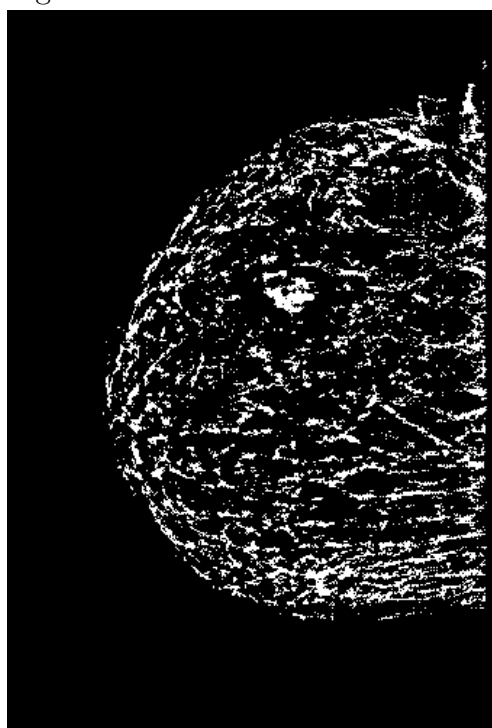


Figura 4.22: Sogliatura centrale, bordo destro e bordo sinistro.

pletamente isolati, che, tra l'altro, non potranno mai essere lesioni tumorali, in quanto agglomerati troppo piccoli.

Per ridurre elementi e migliorare la morfologia delle regioni in generale, l'ultimo passo della fase di pre-detection, è l'applicazione di *Operatori Morfologici*. Verranno introdotti solamente gli operatori di *Apertura*, *Erosione* e *Dilatazione*, in quanto quelli utilizzati nel sistema.

Si definisce *elemento strutturale* (*Structuring Elements*) un pattern definito, specificato come coordinate di un insieme di elementi relativamente ad un origine data. Il modo più comune per rappresentarlo è attraverso una immagine binaria, in rappresentazione matriciale, dove sono in evidenza gli elementi del pattern.

Se consideriamo i valori delle matrici binarie come 1 e 0, un esempio di elemento strutturale circolare è il seguente:

$$\frac{1}{25} * \begin{array}{|c|c|c|c|c|c|c|} \hline 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ \hline 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ \hline \end{array}$$

Il valore in evidenza è il centro di applicazione.

Chiaramente le dimensioni delle matrici sono variabili sulla base dell'effetto che si vuole ottenere.

Il principio di funzionamento di base è lo stesso per ogni operatore. Data una immagine binaria ne viene fatta una scansione attraverso la matrice dell'elemento strutturale considerato. Il passo di traslazione ha valore , in genere, di un pixel, in modo tale che il centro di quest'ultima, nell'arco dell'intero processo, venga a corrispondere ad ogni pixel dell'immagine.

Un ipotetico algoritmo potrebbe essere:

```
for (i=1 to Num_row_image)
  for (j to Num_col_image)
  {
    for (k=(-Dim_mat/2) to (Dim_mat/2))
      for (m=(-Dim_mat/2) to (Dim_mat/2))
      {
        Ni = i+k
        Nj = j+m
        if Ni,Nj inside image
          image[i][j] = image[i][j] OP mat[Ni][Nj]
      }
  }
```

Dove *OP* è specifico dell'operatore che si vuole implementare.

4.5.1 Erosione

È uno degli operatori di base ed ha l'effetto di erodere i bordi degli oggetti, da cui ne consegue che buchi presenti nell'immagine diventano più grandi, mentre le regioni tendono ad assottigliarsi. Una importante conseguenza, che è quella per la quale viene applicato nel CAD, è l'eliminazione di regioni molto piccole.

L'idea che sta alla base è la seguente: ad ogni passo di scansione, il pixel dell'immagine, centrato sulla matrice dell'elemento strutturale, diventa 1, se tale è il valore di ogni altro pixel corrispondente all'elemento strutturale (Vedi Figura 4.23).

Si può notare che il processo eseguito non è altro che la realizzazione di un'operatore logico *AND*. Per cui l'algoritmo sopra indicato diventa, nella sua parte centrale:

```
if Ni,Nj inside image
  image[i][j] = image[i][j] AND mat[Ni][Nj]
```

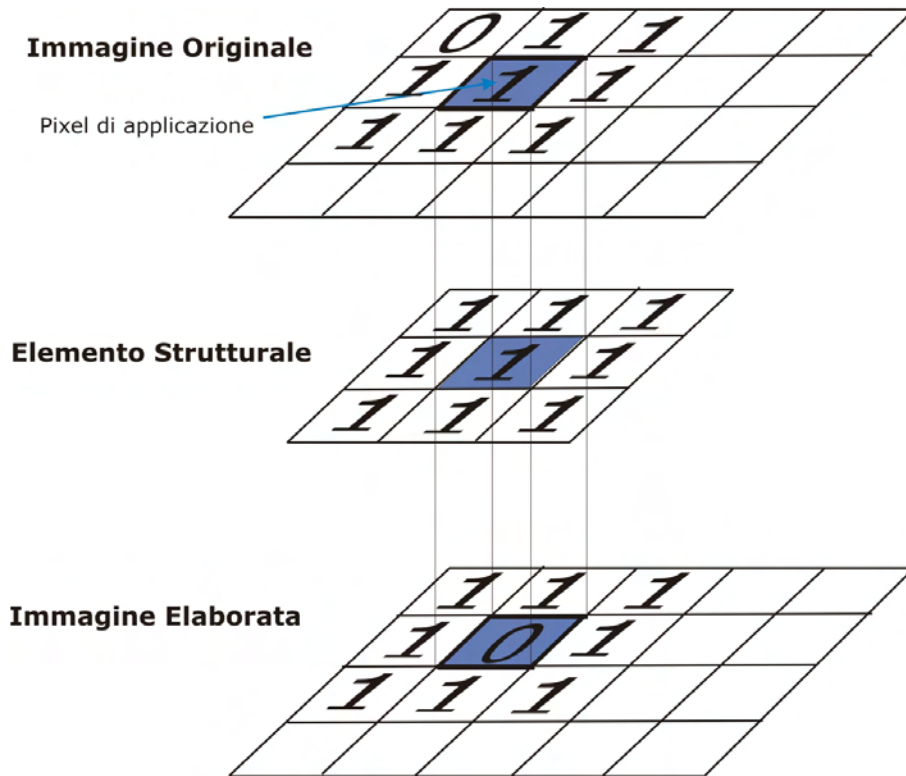


Figura 4.23: Operazione di erosione con elemento strutturale 3x3 quadrato

Formalizzando il problema si ha :

Si definisce \mathbf{X} un insieme di coordinate Euclidee corrispondenti alla immagine binaria di input, e \mathbf{K} l'insieme delle coordinate dell'elemento strutturale. Sia Kx la traslazione di \mathbf{K} tale che la sua origine si in $x \in \mathbf{X}$. L'erosione di \mathbf{X} dato \mathbf{K} è semplicemente l'insieme di tutti i punti x tale che Kx è un sottoinsieme di \mathbf{X} .

Un esempio di applicazione dell'operatore morfologico di erosione è rappresentato in figura 4.27.

4.5.2 Dilatazione

L'altro operatore di base, insieme all'erosione, è la dilatazione, che come dice il nome, ne è l'opposto. L'effetto ottenuto dalla sua applicazione è di allargare i bordi delle regioni, addolcendo eventuali irregolarità.

Essendo l'opposto dell'erosione si avrà che ad ogni passo di scansione il pixel centrato sulla matrice dell'elemento strutturale, diventerà 1 se tale è il valore di almeno un altro pixel corrispondente all'elemento strutturale (Vedi Figura 4.24).

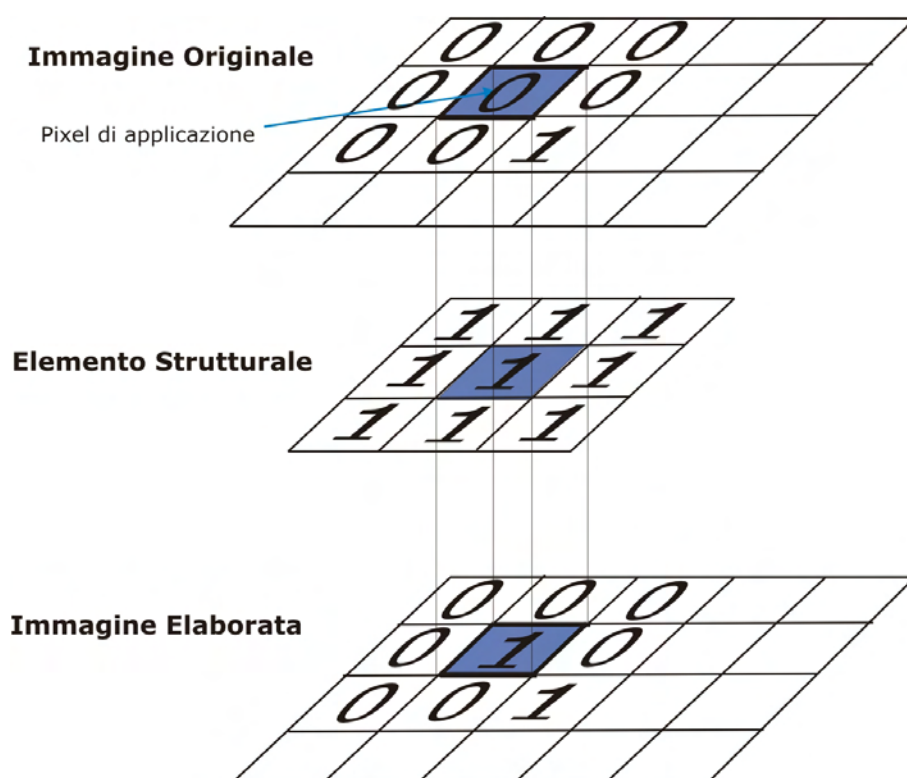


Figura 4.24: Operazione di dilatazione con elemento strutturale 3x3 quadrato

In questo caso l'operazione logica realizzata è un *OR*. Con risultante vari-

azione sull'algoritmo:

```
if Ni,Nj inside image
    image[i][j] = image[i][j] OR mat[Ni][Nj]
```

Formalizzando il problema:

Then the dilation of X by K is simply the set of all points x such that the intersection of Kx with X is non-empty.

Si definisce \mathbf{X} un insieme di coordinate Euclidee corrispondenti alla immagine binaria di input, e \mathbf{K} l'insieme delle coordinate dell'elemento strutturale. Sia Kx la traslazione di \mathbf{K} tale che la sua origine si in $x \in \mathbf{X}$. La dilatazione di \mathbf{X} dato \mathbf{K} è semplicemente l'insieme di tutti i punti x tale che l'intersezione di Kx con \mathbf{X} è non vuota ($Kx \cap \mathbf{X} \neq \emptyset$).

Un esempio di applicazione dell'operatore morfologico di dilatazione è rappresentato in figura 4.26.

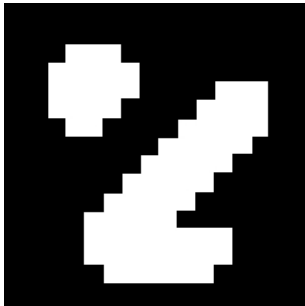


Figura 4.25: Immagine originale.

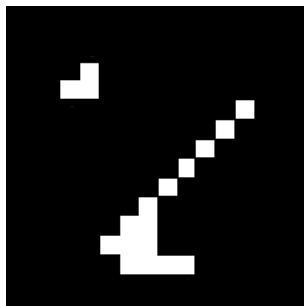


Figura 4.26: ... dopo erosione.

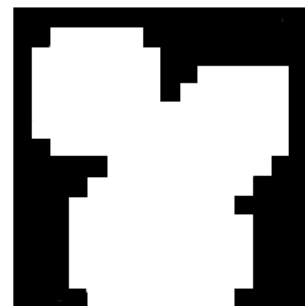


Figura 4.27: ... dopo dilatazione.

4.5.3 Apertura

Viene definita come l'applicazione in cascata di una erosione seguita da una dilatazione.

Per cui l'effetto è quello di avere sì una riduzione delle regioni, ma meno distruttiva rispetto all'erosione, in quanto la dilatazione permette il raffinamento dei bordi.

Applicato all'immagine derivata dalle fase di threshold, ciò che si vuole ottenere è di eliminare pixel isolati o gli agglomerati di piccolissime dimensioni, senza rovinare eccessivamente la forme delle regioni che invece sono di interesse.

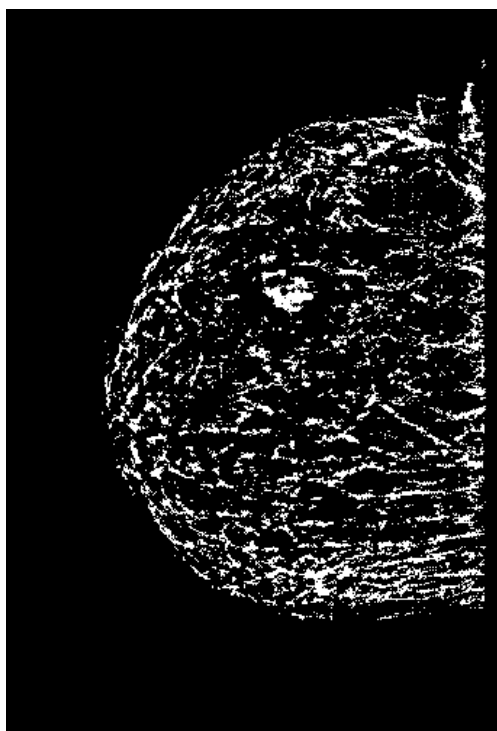


Figura 4.28: Mammogramma generato dalla fase di threshold.

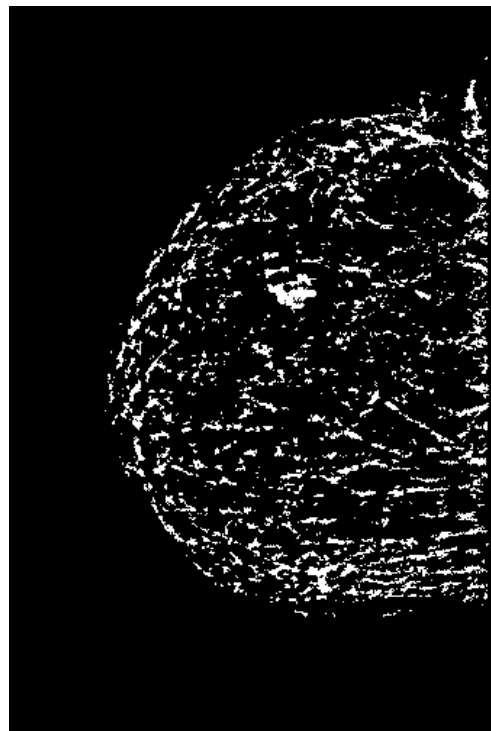


Figura 4.29: Mammogramma dopo l'operazione di Apertura.

Nel sistema CAD presentato si utilizza un elemento strutturale di dimensioni 3x3 quadrato. La scelta è stata guidata da una fase di sperimentazione

nella quale si è posto in evidenza come una matrice anche di poco superiore portava ad un notevole degrado delle prestazioni sul numero di lesioni tumorali individuate.

Si noti in figure ?? e 4.29 con egli operatori morfologici riducano le zone bianche lasciando pressochè inalterate le strutture di dimensione medio grandi.

4.6 Conclusioni

Come ultima fase del processo di pre-detection, l'immagine, rappresentata in figura 4.29, viene riportata alla dimensione del mammogramma originale, attraverso un algoritmo di resize lineare, e passata in input alla fase di pre-elaborazione per l'estrazione dei crop.

Lo scopo, che ci si era prefissati con l'introduzione del modulo, era quello della generazione di un segmentazione nella quale l'area delle regioni di interesse fosse molto inferiore all'intera mammella. Come si vedrà nel capitolo successivo i risultati sono molto promettenti.

Quindi inserendo il Modulo di Pre-detection nel CAD, l'estrazione dei crop non avverrà più facendo un scansione di tutto il mammogramma, ma considerando solamente le regioni individuate dal modulo e questo risulterà molto conveniente, non solo da un punto di vista puramente computazionale.

Capitolo 5

Risultati

5.1 Parametri di valutazione

L'introduzione di un sistema sistema CAD come sostegno al radiologo nella rivelazione di carcinomi all'apparato mammario può portare a risultati di diagnosi estremamente buoni ad una sola condizione: non deve introdurre ulteriore errore. Come si è ampiamente detto le possibilità di errore di un sistema di diagnosi basato solo sulla valutazione umana e da parte di un singolo individuo ha valori fra il 10% e il 30%.

È chiaro che la condizione necessaria per l'introduzione di un ulteriore elemento diagnostico, come il CAD, ha senso nella maniera in cui esso riesce ad abbassare la percentuale di errore.

Si ricorda che vengono considerati errori, non solo le lesioni non individuate, ma anche il numero dei cosiddetti Falsi Positivi.

I parametri che verranno usati per la valutazione del sistema, sono allora i seguenti:

- Percentuale dei Falsi Negativi sul numero totale di lesioni (FN)
- Percentuale dei Falsi Positivi sul numero totale di lesioni (FP)

Un qualsiasi sistema di rivelazione automatico non verrà utilizzato con frequenza se ha dei tempi di elaborazione troppo elevati. Per cui si aggiunge un nuovo parametro di valutazione a quelli già descritti:

- Tempo di risposta del sistema ¹

Quindi la fase di sperimentazione del sistema si è concentrata sull'ottimizzazione di questi tre parametri.

5.2 Database di immagini

Per le varie fasi di prova e ottimizzazione del sistema è stato utilizzato un Database (DDSM ²) di immagini dell'Università della Florida (USF: University of South Florida) appositamente creato per lo screening di massa. Ogni caso è composto dai quattro mammogrammi di analisi, ognuno accompagnato da un file di testo (overlay) di informazione sulla eventuale tipologia di lesione, una sua sommaria classificazione, un indicativo grado di difficoltà, la posizione e il raggio della lesione.

I contorni delle lesioni sono perfezionati.

L'insieme completo è composto di 2500 casi, dai quali sono stati scelti, quelli digitalizzati con scanner Lumisys e Howtek, con rispettive accuratèzze di $50\mu m$ e $43.5\mu m$. Le immagini sono codificate in scale di grigio con valori di luminosità a 12 bit.

Il numero totale di immagini utilizzate ammonta a 1420 mammogrammi, i quali sono stati divisi come mostrato in tabella 5.1.

Nella tabella è stato inserito anche il numero di casi, dove con ciò si identifica il numero di pazienti. Per ognuno di questi si hanno le quattro viste di un normale processo di analisi. Per quanto riguarda quelli maligni sono stati

¹Periodo che intercorre fra il momento in cui viene sottoposta una richiesta al sistema ed una sua risposta.

²Digital Database for Screening Mammography.

	Addestramento		Test		
	Maligne	Sane	Maligne	Benigne	Sane
Crop	996	1500	171	73	$> 10^6$
Mammogrammi	448	655	165	73	79
Casi			110		

Tabella 5.1: Composizione dell'insieme dei mammogrammi utilizzati nelle varie fasi di sperimentazione del CAD.

utilizzati solo i mammogrammi con masse. Considerando che in generale in un paziente si riscontra la patologia solamente in una mammella, per ogni caso si hanno due mammogrammi malati e due sani. Sono possibili alcune eccezioni quali, pazienti con entrambe le mammelle malate, che vengono considerate entrambe appartenenti allo stesso caso, oppure analisi nelle quali il carcinoma è visibile esclusivamente in una vista. In quest'ultima situazione il caso è composto da un solo mammogramma.

Esiste anche la possibilità che vi siano più lesioni in ogni vista.

I risultati del CAD verranno presentati per caso. È infatti sufficiente sapere se una paziente è malata. A tale scopo si ricorda che il responso del CAD non vuole e non deve essere esaustivo, ma consultivo.

Per consistenza dei due insiemi, addestramento e test, i mammogrammi sono stati distribuiti equamente per grado di difficoltà.

5.3 Metodologie di valutazione

L'obiettivo prefissato è quello di ottimizzare le prestazioni del sistema CAD, sia da un punto di vista medico, che computazionale.

La sperimentazione è stata eseguita in due fasi successive:

- ottimizzazione dei parametri del modulo di pre-detection

- ottimizzazione dei parametri del CAD con successivo inserimento del modulo di pre-detection.

È doveroso puntualizzare che il sistema CAD è un progetto in via di sviluppo già da alcuni anni, per cui la seconda fase sopra indicata è il sunto dell'esperienza e dei risultati maturati lungo tutto il periodo.

5.3.1 Pre-detection

L'introduzione del modulo di pre-detection è derivato, in primo luogo, dalla consapevolezza che la sua applicazione avrebbe portato alla riduzione dell'area totale delle regioni di interesse, necessarie nella successiva fase di estrazione dei crop. Quindi ha giocato un ruolo fondamentale l'aspetto computazionale.

Essendo questa un'applicazione medica, che tratta patologie particolarmente difficili da individuare e, soprattutto molto delicate da un punto di vista psicologico del paziente, ci si è posti un limite inferiore al miglioramento delle prestazioni, sulla base del numero di falsi negativi³ accettabili. Tale limite è quello di avere una percentuale di lesioni perse inferiore al 2% sul totale: una regione della mammella contenente una massa, scartata in questa fase, non verrà mai passata al classificatore, e quindi mai identificata.

I parametri di interesse per la valutazione delle prestazioni, per quanto riguarda la pre-detection, sono:

- riduzione dell'area totale, direttamente connessa con l'aumento della velocità di elaborazione dell'intero sistema CAD.
- percentuale di lesioni tumorali scartate in quanto non presenti nella segmentazione apportata, rispetto al numero totale.

La metodologia di conteggio dei Falsi Negativi è la seguente: dato il Ground Truth del carcinoma, o dei carcinomi, si considera un quadrato, posizionato al

³Lesioni tumorali non mantenute dal sistema, ma perse.

centro di essa, di area un nono di quella della massa. Una lesione viene considerata mantenuta dal sistema se esiste una regione, nella nuova segmentazione, con sovrapposizione al quadrato di almeno il 5% dell'area di quest'ultimo.

L'ottimizzazione è avvenuta in diverse fasi, coinvolgendo i seguenti parametri:

- *Fattore di scala per il filtro passa basso Gaussiano*
- *Fattore di scala per l'High Boost Filter*
- *Fattore α di moltiplicazione per la deviazione standard*
- *Raggio ρ della maschera di Threshold*
- *Fattori κ_L , κ_C e κ_R di moltiplicazione della media, per i bordi e il centro della mammella e relative larghezze della banda r_L , r_C e r_R .*

Le prove sono state realizzate su di un insieme totale di 238 mammogrammi così divise:

- 165 mammogrammi con lesioni maligne, con un totale di 171 masse
- 73 mammogrammi con lesioni benigne, con un totale di 73 masse

Il numero totale di masse è di 244.

Fattori di scala

In una prima fase di sperimentazione si è andati ad analizzare la dipendenza che intercorre fra i due fattori di scala, al variare di α , con i parametri usati per la valutazione del sistema, utilizzando una tecnica Grid Search.

Considerando che le masse di interesse hanno dimensione che varia fra 3mm e 35mm, con punte fino a 50mm in casi particolari, si è fissato il raggio della maschera di threshold ad un valore medio di 25mm. Dopo di chè, si è eseguito il sistema variando i due fattori di scala e il fattore α di moltiplicazione per la deviazione standard.

La tabella A.1 in Appendice riassume i risultati ottenuti.

È utile notare nei dati riportato come vi sia una dipendenza pressochè lineare fra la media dell'area delle regioni rimaste e il fattore α . La stessa dipendenza la si ha anche aumentando i fattori scala. Come già detto nelle sezioni dedicate, la diminuzione di scala di un mammogramma implica un certo appiattimento dei picchi presenti in essa. Quindi la tendenza è di avere regioni più grandi e più uniformi. Inoltre osservando una immagine sulla quale è stato applicato un dimensionamento molto basso si notano molte regioni ma piccole a differenza di una con fattori di scala molto alti nella quale sono presenti poche immagini molto grandi.

Sulla base di queste riflessioni si è scelti alcuni valori dei parametri considerati, per il proseguimento delle prove, sulla base dei dati, ma anche attraverso un'analisi visuale di alcuni campioni di mammogrammi.

Le configurazioni scelte sono riportate di seguito:

- Numero Falsi Negativi su un totale di 244 masse

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	5	3	6	7	15	21
3	4	4	9	11	16	25

- Percentuale di Area della mammella, rimasta dopo la segmentazione di pre-detection

Scala Gauss	Scala Highboost	Fattore α				
2	5	23,4	19,8	16,4	13,4	10,9
3	4	23,2	19,4	16,0	13,0	10,4

Fattore α e raggio della maschera di Threshold ρ

Considerati i fattori di scala risultati ottimi, il passo successivo è quello di trovare i valori di α e ρ , che siano ideali rispetto ai parametri di valutazione. È stata quindi eseguita una ulteriore sessione di valutazione di tipo Grid Search su questi due parametri, mantenendo i fattori di scala già trovati.

Dalla sperimentazione è risultato che il valore ideale per il fattore di moltiplicazione della deviazione standard α è 0.6 con ρ a 5mm.

I valori di FN e Area ottenuti sono i seguenti:

FN			Area		
Cancer	Benign	Totale	Cancer	Benign	Totale
2	0	2	18,7	15,4	17,1

Come mostrato in tabella, per quanto riguarda il numero dei Falsi Negativi, si hanno prestazioni già molto elevate. La percentuale è, infatti, inferiore all'1%. Migliore delle aspettative poste all'inizio della sperimentazione.

Le ottimizzazioni successive sono rivolte solo ed esclusivamente alla diminuzione dell'area mantenuta dalla segmentazione.

In figure 5.1 ed 5.2 vengono presentate le due masse perse dalla segmentazione fatta.

Come si nota sono tipologie di massa totalmente differenti, ma in entrambi il problema è intrinseco nei limiti stessi della metodologia di threshold utilizzata e non nei parametri trovati. La prima è una massa il cui nucleo è molto nascosto rispetto ai bordi, quindi, come si vede dalla segmentazione ciò che rimane è proprio il contorno. La seconda invece è localizzata parzialmente in una zona con valori di luminosità molto più elevati rispetto ai propri, per cui nella fase di threshold la media e la deviazione standard sono molto alti.

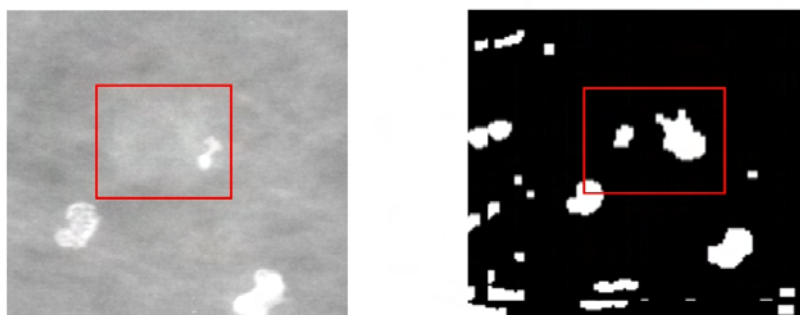


Figura 5.1: Ritaglio di una massa persa dalla segmentazione e segmentazione della stessa.



Figura 5.2: Ritaglio di una massa persa dalla segmentazione e segmentazione della stessa.

Differenziazione Soglia nei bordi

L'idea di permettere la differenziazione della soglia di threshold su tre distinte zone nella mammella, è derivata da un'attenta analisi visuale a posteriori.

Sono diversi i parametri che possono intercorrere al miglioramento di questa fase (vedi paragrafo 4.4):

- larghezza della banda per il bordo interno, vicino al muscolo mammario (da ora r_1), e relativo fattore di moltiplicazione della media (da ora κ_1)
- larghezza della banda per il bordo esterno (da ora r_1), e relativo fattore di moltiplicazione della media (da ora κ_1)

Questi sono tutti parametri che possono essere assegnati esternamente dall'utente.

Per come sono stati usati nel sistema, le sogliature sulle bande esterne hanno due funzionalità decisamente differenti. Si è notato che vicino all'attaccatura al busto del paziente si rivela, in genere, una zona abbastanza omogenea, con conseguente valore di soglia molto basso, come già visto nel paragrafo 4.4. Inoltre sono presenti molto spesso imperfezioni dovute al campionamento, rappresentate da una sottile banda totalmente bianca. Il primo scopo di questa fase è l'eliminazione di quest'ultima.

Sono state eseguite diverse prove con verifica visuale sull'immagine, su un insieme ristretto a 15 mammogrammi e con valori di κ_R e r_R crescenti. Una volta terminato il processo si è passati a sperimentare il sistema con i nuovi parametri per vedere un eventuale degrado sul numero di FN. Appurato che ciò non è avvenuto si è applicato lo stesso processo al bordo esterno della mammella.

Qui lo scopo è diverso: non si devono eliminare radicalmente delle parti, ma solo diminuire l'eccessiva concentrazione di regioni segmentate. Per cui l'idea è di usare un valore di κ_L molto piccolo, con r_R grande. Quest'ultimo è sempre stato scelto in modo empirico con verifica visuale. Mentre il primo provando diversi valori progressivi, fino alla degradazione del sistema, su tutte le 238

immagini.

I parametri Ottenuti sono i seguenti:

κ_L	r_L	r_C	κ_R	r_R
0,1	35mm	0	9	5mm

Le prestazioni del modulo di pre-detection sono migliorate moltissimo:

FN			Area		
Cancer	Benign	Totale	Cancer	Benign	Totale
2	0	2	14,1	12,5	13,3

Si noti che con questa ottimizzazione l'area è notevolmente diminuita, pur mantenendo la stessa efficienza sul numero dei Falsi Negativi.

5.3.2 CAD

La sperimentazione del sistema CAD, escluso del modulo dei pre-detection, è in via di sviluppo da diversi anni. Per cui in questa fase di esposizione dei risultati non verranno motivate scelte su parametri che non siano direttamente dipendenti con l'introduzione del nuovo modulo.

Prestazioni da un punto di vista medico

Le prove sono state seguite in tre fasi:

- CAD singolo (da ora CAD A)⁴senza modulo di pre-detection
- CAD singolo (da ora CAD B)con modulo di pre-detection senza ottimizzazione sui bordi della mammella

⁴Senza l'utilizzo di Unione di Classificatori.

- CAD singolo (da ora CAD C) con modulo di pre-detection con ottimizzazione sui bordi della mammella

Si è scelto di non utilizzare l'opzione dell'Unione dei Classificatori, per avere un riscontro più diretto sugli effettivi miglioramenti del sistema.

I parametri di base per tutti e tre i casi sono i seguenti:

- estrazione dei crop avvenuta attraverso scan con maschere di 8, 10, 15, 17, 22, 27, 33 e 40mm, in quanto le lesioni di interesse hanno dimensione fino a 35mm.
- Kernel Polinomiale di livello 2
- Livelli di decomposizione delle Trasformate Wavelet considerati per l'estrazione delle features: 4 e 6
- Numero totale di feature estratte: 2955

Inoltre tutti sono stati addestrati per la rivelazione di masse nucleate, in quanto quelle senza nucleo sono difficilmente caratterizzabili.

IL CAD A è stato addestrato su 536 positivi estratti mammogrammi con lesioni nucleate e 1600 negativi scelti con bootstrap fra 70000, estratti da mammogrammi sani. Gli insiemi di training sono invece comuni fra il CAD B e il CAD C, con intersezione vuota rispetto quelli appena menzionati, composti da 896 positivi estratti da mammogrammi con masse nucleate e 1500 negativi estratti da mammogrammi sani. In entrambe le tipologie le immagini sono state segmentate dal modulo di pre-detection.

Nella fase di addestramento le parti di pre-detection di ottimizzazione sui bordi sono state disabilitate, in quanto non è di interesse l'area, ma la massima efficienza sui Veri Positivi.

I test sono stati eseguiti facendo eseguire i CAD sui seguenti insiemi:

- 110 casi fra maligni e benigni per determinare il numero dei Veri Positivi

- 79 casi sani per determinare il numero dei Falsi Positivi

I risultati dei tre sistemi a confronto sono rappresentati in tabella 5.2.

CAD A		CAD B		CAD C	
VP	FP	VP	FP	VP	FP
90 (82%)	174 (2,2)	90 (82%)	171 (2,1)	90 (82%)	128 (1,6%)

Tabella 5.2: Risultati dei tre CAD applicati agli stessi insiemi di test. I valori fra parentesi nelle colonne dei Falsi Positivi indicano in numero di Falsi Positivi per immagine.

Come si può notare, l'utilizzo del modulo di pre-detection porta in generale alcuni miglioramenti e si ha un notevole abbassamento del numero di falsi positivi utilizzando l'ottimizzazione sui bordi. Questo è dovuto dal fatto che sono le zone più complesse da un punto di vista di classificazione, per cui maggiore selezione a priori abbassa la possibilità di errore dell'SVM.

Prestazioni da un punto di vista computazionale

L'introduzione di una segmentazione interna ha portato, come abbiamo visto, ad una diminuzione dell'area dell'83%. È indubbio il vantaggio da un punto di vista computazionale che se ne trae.

Sono state fatte alcune prove per verificare il miglioramento della velocità di esecuzione preventivato.

Le specifiche del sistema sono le seguenti: è stato utilizzato un cluster composto da quattro elaboratori con a bordo 2 processori AMD Athlon con frequenza di clock a 1533 Mhz.

I tabella 5.3 sono riportati i tempi di computazione dei vari cad analizzati. I tempi sono riferiti ad un campione casuale di 100 immagini.

Si nota immediatamente l'enorme miglioramento ottenuto. Analizzando i dati si vede come vi sia una dipendenza diretta fra l'area della segmentazione e il tempo di computazione: una riduzione dell'area dell'83% ha portato ad un abbassamento dei tempi dell'83% circa.

CAD A	CAD B	CAD C
1582	316 + 4,2	205 + 4,2

Tabella 5.3: Confronto fra i tempi di computazione dei tre CAD. I tempi sono espressi in secondi e riferiti un campione casuale di 100 immagini. I valori aggiunti sono i tempi del modulo di pre-detection.

Output del CAD

In conclusione si vuole presentare alcuni esempi di mammogrammi processati da un sistema CAD. I riquadri rossi sono le zone classificate positive. Quelle all'interno del Ground Truth segnato in violetto sono i Veri Positivi, gli altri sono Falsi Positivi.

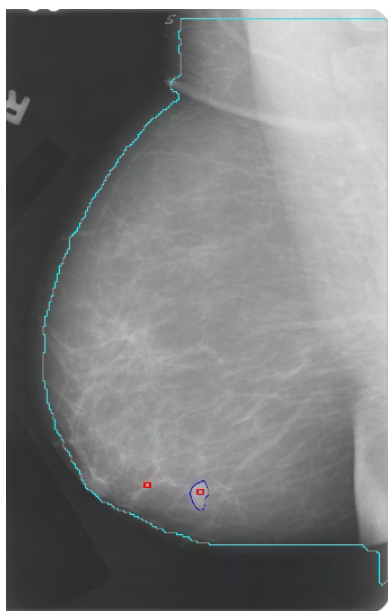


Figura 5.3: Mammogramma classificato con 1 VP ed 1 FP.

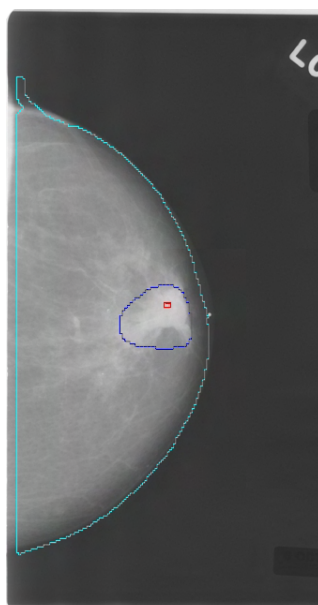


Figura 5.4: Mammogramma classificato con 1 VP ed 0 FP.

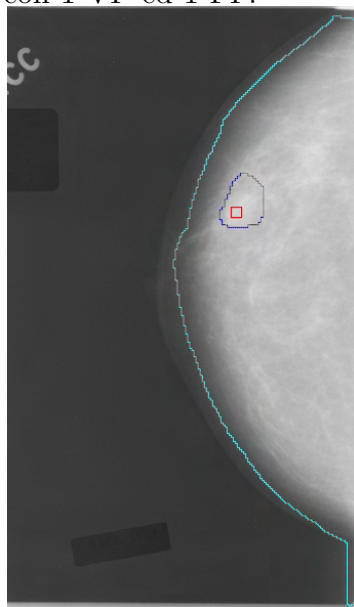


Figura 5.5: Mammogramma classificato con 1 VP ed 1 FP.

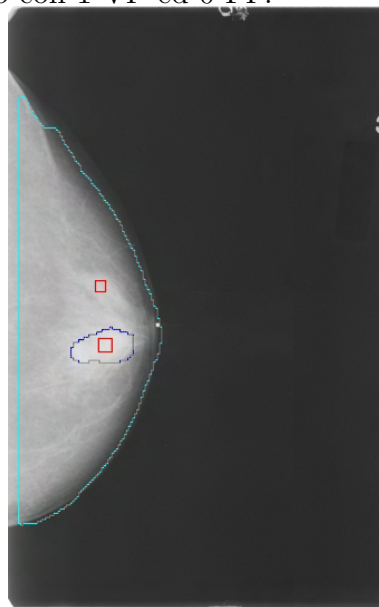


Figura 5.6: Mammogramma classificato con 1 VP ed 1 FP.

Capitolo 6

Conclusioni e Sviluppi Futuri

La rivelazione di lesioni tumorali al seno è un compito alquanto complesso. Le percentuali di errore da parte dell'uomo sono ancora troppo alte. Affiancare in veste consultiva un sistema CAD può essere una soluzione. Chiaramente ciò ha senso se esso non introduce ulteriore errore. I risultati ottenuti confermano l'efficacia del metodo.

La vera innovazione è stata l'introduzione del modulo di pre-detection nel cad per eseguire una segmentazione interna, e si sono ottenuti risultati assolutamente inaspettati, soprattutto da un punto di vista medico. I miglioramenti sulla velocità computazionale erano abbastanza scontati, gli altri invece no. Non si pensava ad una così drastica diminuzione dei Falsi Positivi. Quindi da ora in poi lo sviluppo del CAD, come sistema globale, punterà molto anche sulla fase di pre-detection.

Nella descrizione del modulo sono state evidenziate alcune problematiche che esistono. Quindi i lavori futuri verteranno sul risolvere queste. In primo luogo superare le limitazioni che ha. Si è visto come esso non riesca bene nell'individuazione di alcuni tipi di masse. Un possibile miglioramento potrebbe essere quello di utilizzare il threshold, come fatto fino ad ora, e poi segmentare meglio le regioni attraverso algoritmi di region growing.

Un'altra considerazione è la seguente: il modulo è molto dipendente, per sua logica, alla geometrie e morfologie dell'immagine analizzata. I parametri sono stati ottimizzati per le tipologie di mammogrammi a nostra disposizione. Se queste cambiano come reagisce il sistema? Uno sviluppo futuro è quello di verificare la robustezza del sistema e cercare di generalizzare il più possibile. Questo soprattutto in vista della mammografia digitale.

Il vero futuro del connubio fra Mammografia e Informatica è proprio nella mammografia digitale.

Appendice A

Tabelle - Parametri Grid Searching

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	43	55	69	88	100
2	3	17	23	34	45	56
2	4	4	12	16	22	30
2	5	2	4	5	9	15
2	6	2	4	7	9	10
3	2	22	29	41	52	57
3	3	5	11	20	28	31
3	4	2	5	6	9	17
3	5	3	7	8	10	11
3	6	3	4	4	7	10
4	2	9	11	24	32	41
4	3	5	6	8	16	23
4	4	5	7	10	13	15
4	5	4	4	5	7	13
4	6	14	16	16	21	28

Tabella A.1: Masse Perse su un totale di 171 - Cancer

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	13	21	28	33	46
2	3	7	10	12	14	18
2	4	2	5	6	10	14
2	5	1	2	2	6	6
2	6	1	2	3	4	6
3	2	6	9	13	17	22
3	3	2	5	6	8	16
3	4	2	4	5	7	8
3	5	1	3	5	6	8
3	6	3	3	5	7	9
4	2	4	4	8	11	18
4	3	2	4	7	8	9
4	4	9	4	6	7	10
4	5	3	4	4	6	9
4	6	4	5	8	9	10

Tabella A.2: Masse Perse su un totale di 73 - Benign

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	56	76	97	121	146
2	3	24	33	46	59	74
2	4	6	17	22	32	44
2	5	3	6	7	15	21
2	6	3	6	10	13	16
3	2	28	38	54	69	79
3	3	7	16	26	36	47
3	4	4	9	11	16	25
3	5	4	10	13	16	19
3	6	6	7	9	14	19
4	2	13	15	32	43	59
4	3	7	10	15	24	32
4	4	14	11	16	20	25
4	5	7	8	9	13	22
4	6	18	21	24	30	38

Tabella A.3: Masse Perse su un totale di 244 - Totale Cencer + Benign

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	0,094	0,075	0,059	0,047	0,038
2	3	0,130	0,107	0,087	0,072	0,059
2	4	0,180	0,151	0,126	0,104	0,086
2	5	0,257	0,217	0,181	0,149	0,120
2	6	0,355	0,300	0,247	0,196	0,150
3	2	0,110	0,089	0,072	0,058	0,046
3	3	0,171	0,142	0,118	0,097	0,079
3	4	0,254	0,214	0,177	0,144	0,115
3	5	0,353	0,298	0,244	0,193	0,147
3	6	0,449	0,372	0,292	0,214	0,146
4	2	0,149	0,122	0,099	0,079	0,063
4	3	0,245	0,204	0,168	0,135	0,106
4	4	0,349	0,293	0,238	0,186	0,139
4	5	0,447	0,368	0,285	0,205	0,136
4	6	0,448	0,363	0,270	0,183	0,112

Tabella A.4: Media dell'area mantenuta all'interno della segmentazione - Cancer

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	0,084	0,067	0,053	0,042	0,034
2	3	0,113	0,092	0,076	0,062	0,051
2	4	0,150	0,125	0,103	0,085	0,071
2	5	0,213	0,178	0,147	0,120	0,097
2	6	0,301	0,251	0,204	0,162	0,125
3	2	0,097	0,078	0,062	0,050	0,040
3	3	0,142	0,117	0,096	0,078	0,064
3	4	0,210	0,174	0,143	0,116	0,093
3	5	0,300	0,249	0,202	0,159	0,122
3	6	0,403	0,331	0,259	0,192	0,134
4	2	0,118	0,095	0,075	0,059	0,046
4	3	0,198	0,163	0,132	0,105	0,082
4	4	0,293	0,242	0,194	0,151	0,113
4	5	0,399	0,325	0,251	0,182	0,124
4	6	0,426	0,339	0,250	0,168	0,104

Tabella A.5: Media dell'area mantenuta all'interno della segmentazione - Benign

Scala Gauss	Scala Highboost	Fattore α				
2^n	2^m	0,20	0,40	0,60	0,80	1,00
2	2	0,089	0,071	0,056	0,045	0,036
2	3	0,121	0,100	0,082	0,067	0,055
2	4	0,165	0,138	0,114	0,095	0,078
2	5	0,234	0,198	0,164	0,134	0,109
2	6	0,328	0,275	0,225	0,179	0,138
3	2	0,104	0,083	0,067	0,054	0,043
3	3	0,157	0,130	0,107	0,088	0,071
3	4	0,232	0,194	0,160	0,130	0,104
3	5	0,327	0,273	0,223	0,176	0,134
3	6	0,426	0,351	0,275	0,203	0,140
4	2	0,134	0,108	0,087	0,069	0,055
4	3	0,221	0,184	0,150	0,120	0,094
4	4	0,321	0,267	0,216	0,169	0,126
4	5	0,423	0,346	0,268	0,194	0,130
4	6	0,437	0,351	0,260	0,175	0,108

Tabella A.6: Media dell'area mantenuta all'interno della segmentazione - Cancer + Benign

Ringraziamenti

In conclusione al lavoro è doveroso un ringraziamento alle persone che in un modo o nell'altro mi sono state sempre vicine. In primo luogo al Prof. Campanini per la grande disponibilità dimostrata ad aiutarmi ad entrare in un mondo affascinante ma a me quasi sconosciuto.

Una menzione particolare al Dott. Roffilli per tutto l'aiuto pratico e per avermi spronato nei momenti in cui c'era bisogno di dare il massimo.

Un grazie a tutti i miei compagni di viaggio del laboratorio di Fisica Sanitaria, per i consigli e il bellissimo clima creato, e a agli informatici che direttamente o indirettamente mi hanno aiutato. In particolare a Fabrizio Bisi e Cesare Quadalti.

Per ultima la mia famiglia che mi ha sempre sostenuto lungo questo viaggio, i miei zii, i miei amici e i gen senza i quali non avrei potuto fare ciò che ho fatto. Fra questi un grazie particolare a Fabio Teofani, Christian Visani e i miei cugini che sono stati e sono per me più che fratelli: Lidia e Marco Lullo.

Bibliografia

- [1] R. Campanini, A. Bazzani *et al*
A novel approach to mass detection in digital mammography based on Support Vector Machine (SVM)
Digital Mammography: IWDM2002, 6th International Workshop on Digital Mammography ed H O Peitgen (Springer) pp 399-401.
- [2] C. Burges
A Tutorial on Support Vector Machine for Pattern Recognition
Bell Lab.
- [3] N. Cristianini, J. Shawe-Taylor
An Introduction to Support Vector Machines
Cambridge University Press 2000.
- [4] I. El-Naqa Y. Yang *et al*
A Support Vector Machine Approach for Detection of Microcalcifications
IEEE Transaction on Medical Imaging, Vol.21, No. 12, December 2002.
- [5] D. L. Pham, C. Xu, J. L. Prince
A Survey of Current Method in Medical Image Segmentation
Annual Review of Biomedical Engineering, 1998.
- [6] W. Tao, H. Burkhardt
An Effective Image Thresholding Method Using a Fuzzy Compactness Measure
12th Conference on Pattern Recognition, Israel, October 1994.

-
- [7] , H. Burkhardt
Adaptive Image Region-Growing
IEEE Transaction on Image Processing, Vol. 3, No. 6, November 1994.
- [8] N. Petrick, B. Sahiner, H. P. Chan *et al*
Breast Cancer Detection: Evaluation of a Mass-Detection Algorithm for Computer-aided Diagnosis
- [9] A. Riccardi
Clasificación de imágenes mamográficas con el método SVM
tesi di laurea, Università degli studi di Bologna, Corso di laurea in Fisica, AA 1999/2000.
- [10] J. Roebuck
Clinical Radiology of the Breast
Heinemann Medical Books, Oxford, 1990
- [11] J. Ferlay, F. Bray, P. Pisani and D.M. Parkin
Computer Aided Detection in Mammography
Working Party of the radiologists Quality Assurance Coordinating Group, NHSBSP publication N48, January 2001
- [12] Sito Web *DDSM - Digital Database for Screening Mammography* California University - URL:
<http://marathon.csee.usf.edu/Mammography/Database.html>
- [13] R. C. Gonzales, R. E. Woods
Digital Image Processing
Addison-Wesley Publishing Company, 1993
- [14] K. R. Castelman
Digital Image Processing
Practice Hall, Englewood Cliff, New Jersey 1996
- [15] B. E. Henderson, M. C. Pike e R. K. Ross
Epidemiology and Risk Factor

- Bonadonna (ed) Breast Cancer: Diagnosis and Management. John Wiley & Sons, 15-33, 1984
- [16] J. Ferlay, F. Bray, P. Pisani and D.M. Parkin
GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence World-wide
Version 1.0. IARC CancerBase No. 5. Lyon, IARCPress, 2001
- [17] I. L. Kuncheva, C. J. Whitake *at al*
Is Indipendence Good for Combining Classifiers?
Proceeding of the 15th International Conference on Pattern Rcongnition, Barcelona, Spain, Vol2 pp 168-71
- [18] M. B. Ahmad, T. C. Choi
Local Threshold and Boolean Function Based Edge Detection
IEEE Transactions on Consumer Electronics, Vol45, No. 3, AUGUST 1999
- [19] R.Fletcher
Prectical Method of Optimization
J. Wiley, 1998
- [20] O. Schiaratura
Progettazione ed Implementazione di un sistema di calcolo ibrido Multithread-Multiprocesso per HPC: applicazione all'imaging medico
tesi di laurea, Università degli studi di Bologna, Corso di laurea in Scienza dell'informazione, AA 2002/2003.
- [21] R. Tazzoli
Rivelazione di Masse in Mammografia Digitale
tesi di laurea, Università degli studi di Bologna, Corso di laurea in Fisica, AA 2000/2001.
- [22] D. N. Dongiovanni
Rivelazione di masse in mammografia mediante Wavelet e Support Vector

Machines

tesi di laurea, Università degli studi di Bologna, Corso di laurea in Fisica, AA 2000/2001.

- [23] J. Ferlay, F. Bray, P. Pisani and D.M. Parkin
The Computer Aided Detection of Abnormalities in Digital Mammograms
PhD Thesis, Dep. of Medical Physics, University of Manchester, 1996
- [24] P. Hopwood e P. Maguire
The Psychological Impact of the Diagnosis and Treatment of Breast Cancer
Manchester Breast Screening Manual. University Dept. Of Medical Illustration, Withington Hospital, Manchester, 1990.
- [25] A. Fournier *at al*
Wavelet and their application on Computer Graphics
1995 note del corso num 26 della conferenza Siggraph'95.
- [26] E. J. Stollintz, T. D. DeRose, D. H. Salesin
Wavelets for Computer Graphics: A Primer
University of Washington.