
Università degli Studi di Padova – Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Biomedica

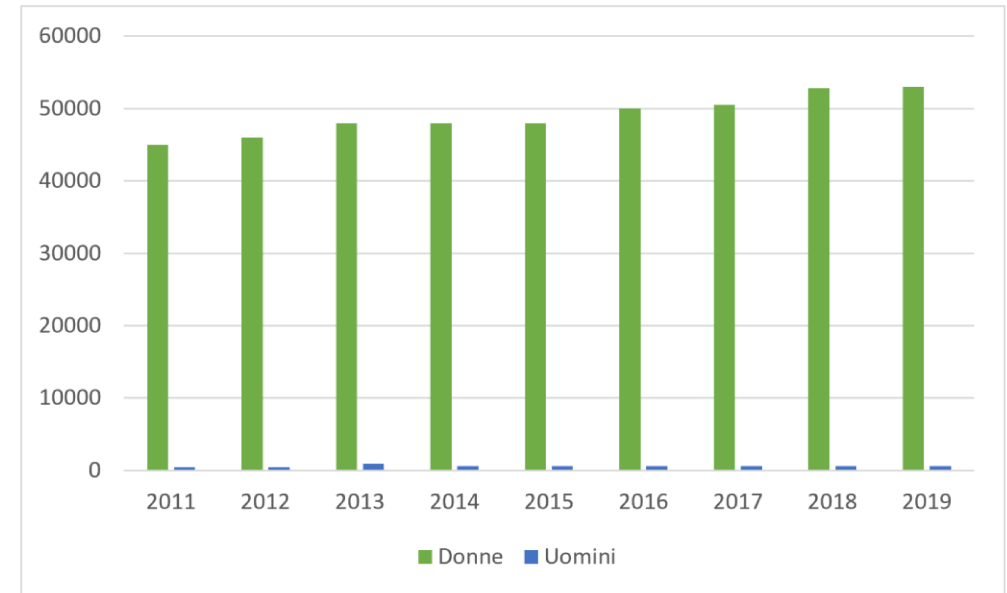
Relazione per la prova finale
***«Applicazione del Machine Learning alla classificazione di
immagini mammografiche: confronto tra algoritmi.»***

Relatore: Prof. Augusto Ferrante

Laureanda: Irene de Florio
Matricola n.: 2152249

Padova, 23/09/2026

- Il tumore al seno costituisce il 10,4% di tutti i casi di tumore a livello globale in percentuali diverse tra uomini e donne (OMS, 2021).
- Tra il 2007 e il 2019 il tasso di mortalità è diminuito del 6% grazie agli screening preventivi e alla ricerca.
- Necessità di tecniche di rilevamento accurate e precise.
- **Obiettivo del lavoro: confronto tra alcuni algoritmi di Machine Learning applicati ad immagini mammografiche per la determinazione della benignità o malignità del tumore.**



Numero di diagnosi di cancro al seno in Italia dal 2011 al 2019 – dati adattati da AIOM

- **Imaging medico:** mammografia a raggi X (utilizzata di routine per lo screening e la diagnosi), ecografia (interpretazione potenzialmente difficile) e risonanza magnetica al seno (per la valutazione pre-intervento chirurgico).
- **Radiomica:** frontiera della medicina che utilizza l'AI basata sull'estrazione di dati quantitativi dalle immagini radiologiche e sul loro uso per la creazione di sistemi di supporto alle decisioni cliniche.
- **Sistemi CAD:** dagli anni '90, progettati per aumentare l'accuratezza dello screening dato dalla MG.
- **Machine Learning (ML):** forma di intelligenza artificiale introdotta negli anni '80. Svolge compiti senza indicazioni di programmazione chiare, identificando associazioni nascoste tra i dati e analizzandole. Nel sistema CAD per l'imaging mammario aiuta nella diagnosi dei tumori al seno e non è influenzato da modalità di lettura del radiologo, stanchezza, distrazione.

FASI DI CLASSIFICAZIONE DI IMMAGINI MAMMOGRAFICHE

- Selezione di un **database** di immagini mammografiche.
- **Feature extraction** (di texture, morfologiche e statistiche) e **feature selection**.
- **Modello** del classificatore: supervisionato, non supervisionato e semisupervisionato.
- **Parametri** di misurazione delle prestazioni

Classe reale	Classe predetta	
	Positiva	Negativa
Positiva	TP	FN
Negativa	FP	TN

$$\text{Sensibilità} = \frac{TP}{TP+FN}$$

$$\text{Precisione} = \frac{TP}{TP+FP}$$

$$\text{Specificità} = \frac{TN}{TN+FP}$$

$$\text{Accuratezza} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Output** del classificatore (benigno o maligno)

- **Logic Based (C4.5)**: seguono il principio della struttura ad albero e dell'argomentazione logica. Hanno le seguenti caratteristiche: root node, splitting, decision node, terminal node, pruning.
- **Support Vector Machine (SVM)**: si basano sulla possibilità di separare i dati in due classi distinte $\{\omega^1, \omega^2\}$ mediante il miglior iperpiano esistente. Sia $X = \{X_i\}$, con $X_i \in \mathbb{R}^n$, $i = \{1, \dots, l\}$. La classificazione di questi dati può essere espressa come $C = (X_1, \omega_1) \dots (X_l, \omega_l)$. Durante la fase di apprendimento trovano un set di parametri $\mathbf{w} = [w^1, w^2, \dots, w^n]^T$ e b per produrre una funzione decisionale d che per ogni i agisce come:

$$d(X_i, \mathbf{w}, b) = \mathbf{w} \cdot X_i + b = \sum_{j=1}^n \mathbf{w}^j X_i^j + b$$

La previsione dell'algoritmo è $X_i \in \omega_1$ se $d \geq 1$ e $X_i \in \omega_2$ se $d < 1$ che corrisponde geometricamente alla bipartizione di $\mathbb{R}^n$ mediante un iperpiano.

- **Bayesian (Naïve Bayes, NB)**: dipende dalla stima delle probabilità, non segue alcuna regola decisionale esplicita e considera ognuna delle caratteristiche statisticamente indipendente. Si basa sulla formula di Bayes:

$$P(C_k | \mathbf{x}^k) = \frac{P(\mathbf{x}^k | C_k)P(C_k)}{P(\mathbf{x}^k)}$$

Dove $\mathbf{x}^k$ è un vettore delle caratteristiche appartenente ad una classe C_k .

Il classificatore si esprime come:

$$\bar{c} = \arg \max_{k \in \{1, \dots, m\}} P(C_k) \prod_{i=1}^n P(x_i^k | C_k)$$

CONFRONTO: EFFICACIA

Criteri di valutazione	Classificatori		
	C4.5	SVM	NB
Tempo di costruzione del modello (s)	0.06	0.07	0.05
Istanze correttamente classificate	665	678	671
Istanze non correttamente classificate	34	21	28
Accuratezza (%)	95.13	97.13	95.99

- **Tempi di costruzione** del modello molto vicini tra loro.
- **Accuratezza** di SVM maggiore di quella di NB e C4.5.
- SVM ha il valore più alto di **istanze classificate correttamente** e il più basso di **classificazioni erranee**.

CONFRONTO: EFFICIENZA

Criteri di valutazione	Classificatori		
	C4.5	SVM	NB
Kappa Statistic (KS)	0.89	0.93	0.91
Errore assoluto medio (MAE)	0.06	0.02	0.03
Radice dell'errore quadratico medio (RMSE)	0.21	0.16	0.19
Errore assoluto relativo (RAE) (%)	14	6.33	8.59
Radice dell'errore quadratico relativo (RRSE) (%)	45	35.58	40.95

- SVM offre le migliori prestazioni complessive: **elevata precisione** (KS) e **ridotto tasso di falsi positivi** (MAE).
- C4.5 presenta il **valore più alto del tasso di errore**.

CONFRONTO: EFFICIENZA

Classificatori	TP	FP	Precisione	Sensibilità	Classe
C4.5	0.95	0.05	0.96	0.95	Benigno
	0.94	0.04	0.91	0.94	Maligno
SVM	0.97	0.03	0.98	0.97	Benigno
	0.96	0.02	0.95	0.96	Maligno
NB	0.95	0.02	0.98	0.95	Benigno
	0.97	0.04	0.91	0.97	Maligno

- SVM valore più alto nella **predizione di TP** per la classe benigna, Naïve Bayes per la classe maligna.
- SVM e NB mostrano massimo valore di **precisione** per la classe benigna e SVM anche per la classe maligna.
- Complessivamente, i valori di **sensibilità** variano tra 95-97% per b-BT e tra 94-97% per m-BT.

- L'aumento di casi di cancro al seno ha reso ancor più urgente e necessaria una **diagnosi veloce, accurata e precisa**. I sistemi CAD che utilizzano algoritmi basati su Machine Learning sono un importante strumento di ausilio medico, poiché possono provvedere con ottimi risultati alla classificazione di immagini mammarie.
- Gli algoritmi ML possono essere di diverse tipologie: **Logic Based, SVM e Bayesian** sono tra i principali. Tramite un confronto basato sulla letteratura scientifica si è potuto evincere come tutti e tre questi classificatori forniscano degli ottimi risultati quando applicati alla determinazione di benignità o malignità del tumore a partire da immagini BC.
- **L'algoritmo SVM** risulta complessivamente il migliore e difatti ad oggi è il più comunemente utilizzato nell'analisi di immagini mammografiche.

-
- [1] Y. Guo et al., <<Machine learning and new insights for breast cancer diagnosis,>> *International Medical Resarch*, 2024.
- [2] Associazione Italiana di Oncologia Medica, *I numeri del cancro in Italia*, 2025. indirizzo: <https://www.aiom.it/i-numeri-del-cancro-in-italia/>.
- [3] Associazione Europa Donna Italia, 2024. indirizzo: <https://www.europadonna.it/2024/01/26/tumore-al-seno-i-numeri-aggiornati/>.
- [4] Associazione AIRC, *Screening per il tumore del seno*. indirizzo: <https://www.airc.it/cancro/prevenzione-tumore/guida-agli-screening/seno-mammografia>.
- [5] A.-A. Nahid e Y. Kong, <<Involvement of Machine Learning for Breast Cancer Image Classification: A Survey,>> *School of Engineering, Macquarie University, Sydney, NSW 2109, Australia*, 2017.
- [6] H. Asria, H. Mousannifb, H. A. Moatassimec e T. Noeld, <<Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,>> *Procedia Computer Science* 83, 2016.
- [7] H. Yu e S. Kim, <<SVM Tutorial: Classification, Regression, and Ranking,>> *researchgate.net*, 2012.