Deep Learning in EPICA

Leonardo Ventura

ISPRO
Istituto per lo studio, la prevenzione
e la rete oncologica

Il contesto

- Valutazione dei percorsi diagnostico terapeutici e assistenziali in termini di:
  - Qualità dei percorsi
  - Accesso alle cure
  - Appropriatezza
  - Compliance ai protocolli

- Calcolo di specifici indicatori attraverso i flussi amministrativi correnti
  - Schede di dimissione ospedaliera
  - Anatomie patologiche
  - Farmaceutica
  - Prestazioni ambulatoriali
  - Esenzioni

Modalità di calcolo degli indicatori

- Modelli gerarchici aggiustati per controllare l'effetto di possibili confondenti
    - Sesso, età, stadio alla diagnosi, volume ospedaliero, deprivazione

- Utilizzo di variabili estratte dai flussi per calcolare gli indicatori
    - Numero di linfonodi asportati
    - Effettuazione del linfonodo sentinella
    - Recettori estrogenici e progestinici

- Estrazione di «queste» informazioni dai referti delle anatomie patologiche

Anatomie patologiche

Data Accettazione [redacted] Data Referto [redacted] AP 909031 | Tipo I Numero 11811

Reparto CLINICA CHIRURGICA I - SEZ. DEGENZA

Notizie CA. MAMMELLA DX Q.E.C.: AMPIA EXERESI + L.A. IN DA

Macroscopia

1-2) LIMITE DI SEZIONE CHIR. VERSO IL CAPEZZOLO. 3-4) LIMITE DI SEZIONE CHIR. SUPERIORE. 5-6) LIMITE DI SEZIONE CHIR. INFERIORE. 7) BX INTRAOPERATORIA LESIONE. 8-10) LESIONE DIAMETRO MM. 9. 11) PARENCHIMA MAMMARIO. 12-16) LINF. ASCELLARI (24). 17) LINF. SOPRA LA VENA ASCELLARE (INV. SEP.) (1).

Diagnosi

CARCINOMA DUTTALE INFILTRANTE VARIETA' CRIBRIFORME G1. INVASIONE VASCOLARE EMATICA/LINFATICA PERITUMORALE ASSENTE. DISTANZA MINIMA TUMORE-MARGINI DI SEZIONE CHIRURGICA MAGGIORE DI 10 MM. NESSUNA PROLIFERAZIONE NEOPLASTICA NEI LINFONODI ASCELLARI (pT1b, pN0, pMx).
PARAMETRI BIOLOGICI:
ER (clone 1D5): POSITIVO 100% INTENSITA' DELLA COLORAZIONE: MARCATA
PgR (clone 1A6): POSITIVO 30% INTENSITA' DELLA COLORAZIONE: MARCATA
Ki67 (clone MIB1): INFERIORE AL 5%

Anatomie patologiche

- Problematica di estrarre le informazioni rilevanti dai referti AP che si presentano sotto forma di testo libero
- In EPICA1 procedemmo alla lettura manuale attraverso il coinvolgimento di 4-5 assistenti sanitarie
- Lavoro lungo e oneroso
- Più di un referto AP per ciascun paziente
- In EPICA1 lavorammo a circa 3400 casi incidenti di tumore alla mammella nell'anno 2016 con associate più di 11000 referti AP
- In EPICA2 lavoreremo ai casi incidenti di mammella e colon retto per gli anni 2017-2018…

ISPRO
Istituto per lo studio, la prevenzione e la rete oncologica

Un approccio diverso

- Opportunità di utilizzare tecniche più avanzate per l'estrazione delle informazioni dai referti AP
- Approccio automatico
- Metodologie di Intelligenza Artificiale-Machine Learning-Deep Learning

Intelligenza artificiale-machine learning e deep learning

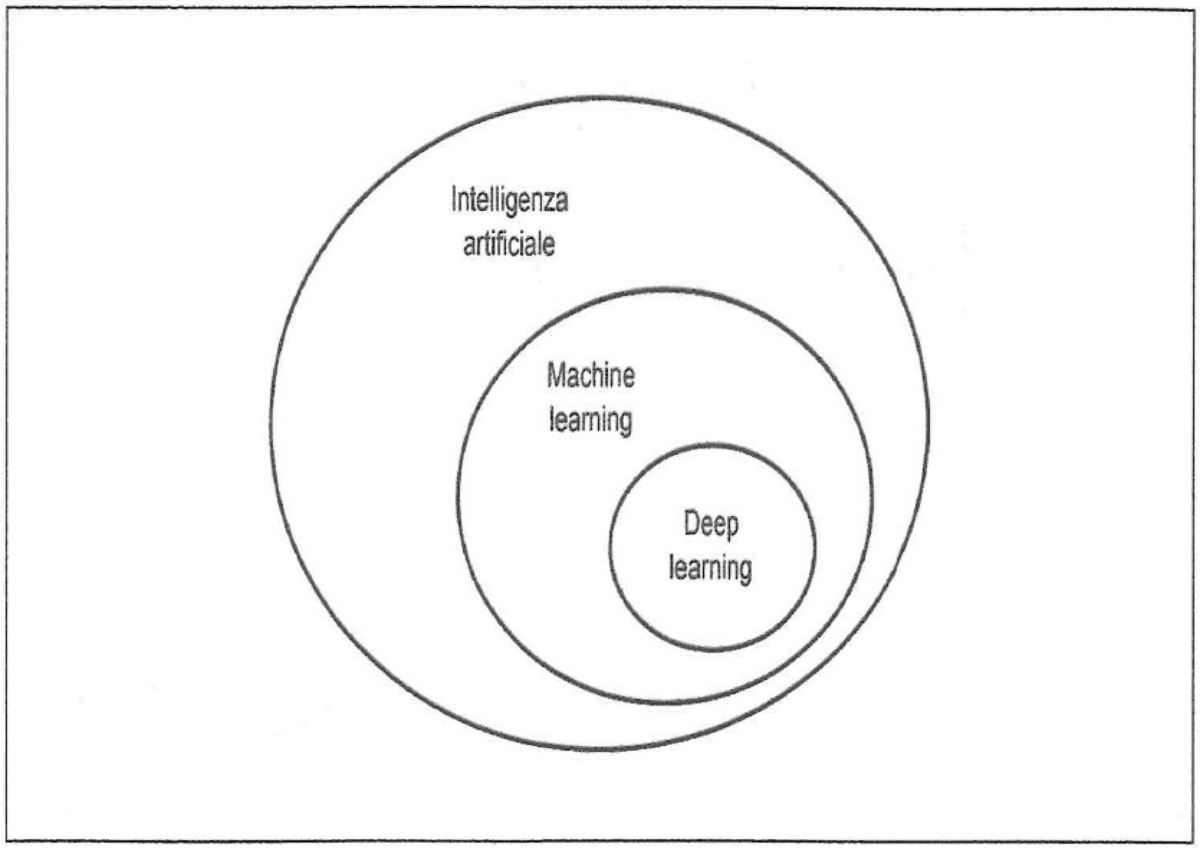


Figura 1.1 Intelligenza artificiale, machine learning e deep learning.

Naïve Bayes Classifier Algorithm
K Means Clustering Algorithm
Support Vector Machine Algorithm
Apriori Algorithm
Linear Regression
Logistic Regression
Artificial Neural Networks
Random Forests
Decision Trees
Nearest Neighbours

Cosa rende «deep» il deep learning?

- Apprendimento delle rappresentazioni (o codifiche) dei dati che pone l'accento su layer (livelli) successivi.
- Si basa sull'idea di *raffinare* layer sempre più «profondi» nelle rappresentazioni.
- Il numero di layer che costituiscono un modello è detto «profondità» del modello.
- Nel riconoscimento delle immagini, queste vengono trasformate in rappresentazioni sempre più differenti rispetto all'immagine originale e sempre più informative sul risultato finale.
- L'immagine attraversa sempre più filtri e ne esce sempre più purificata, e quindi utile a servire un determinato compito.

Rappresentazione profonda

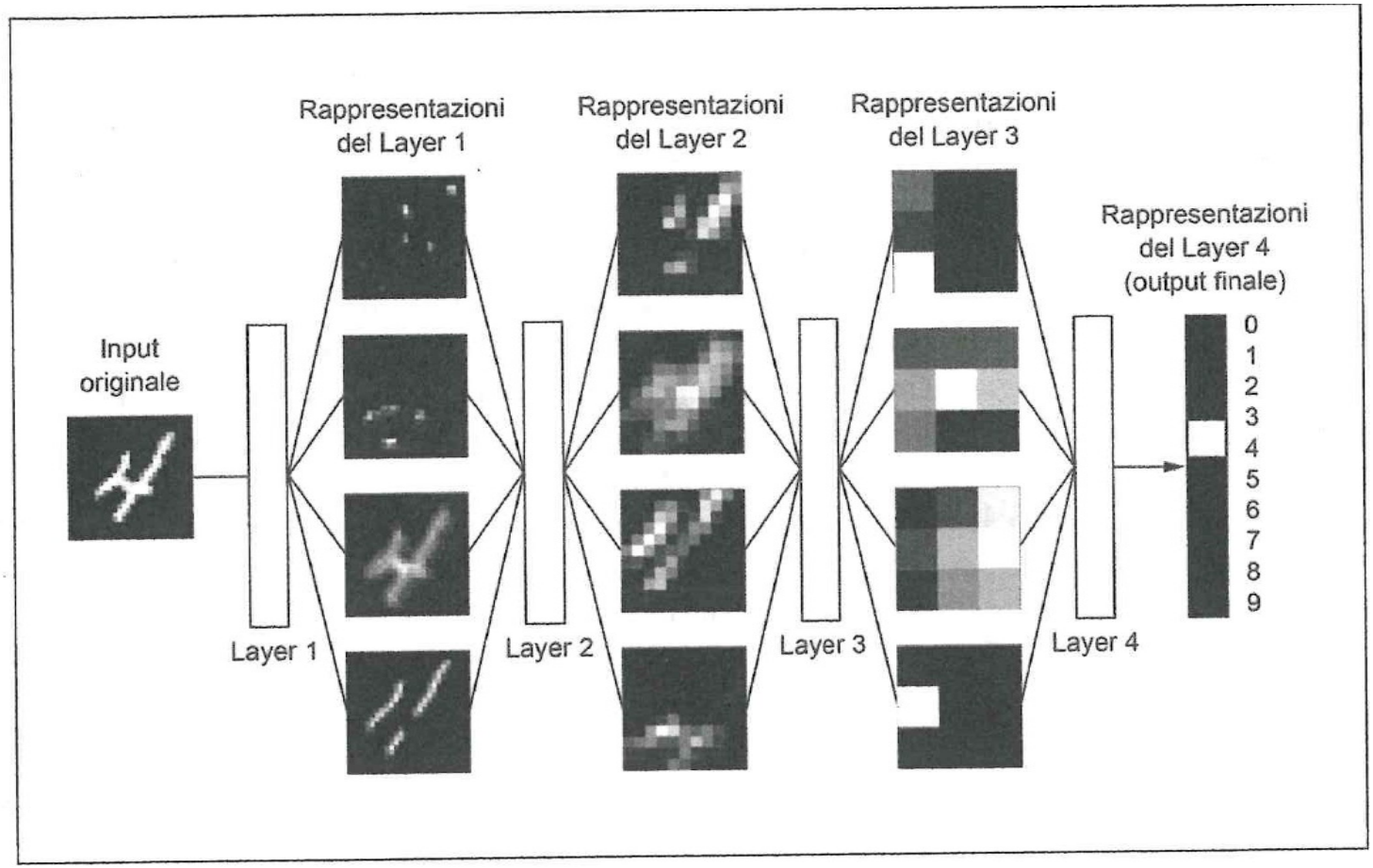


Figura 1.6 Rappresentazioni profonde apprese in base a un modello di classificazione per cifre.

Anatomia di una rete neurale

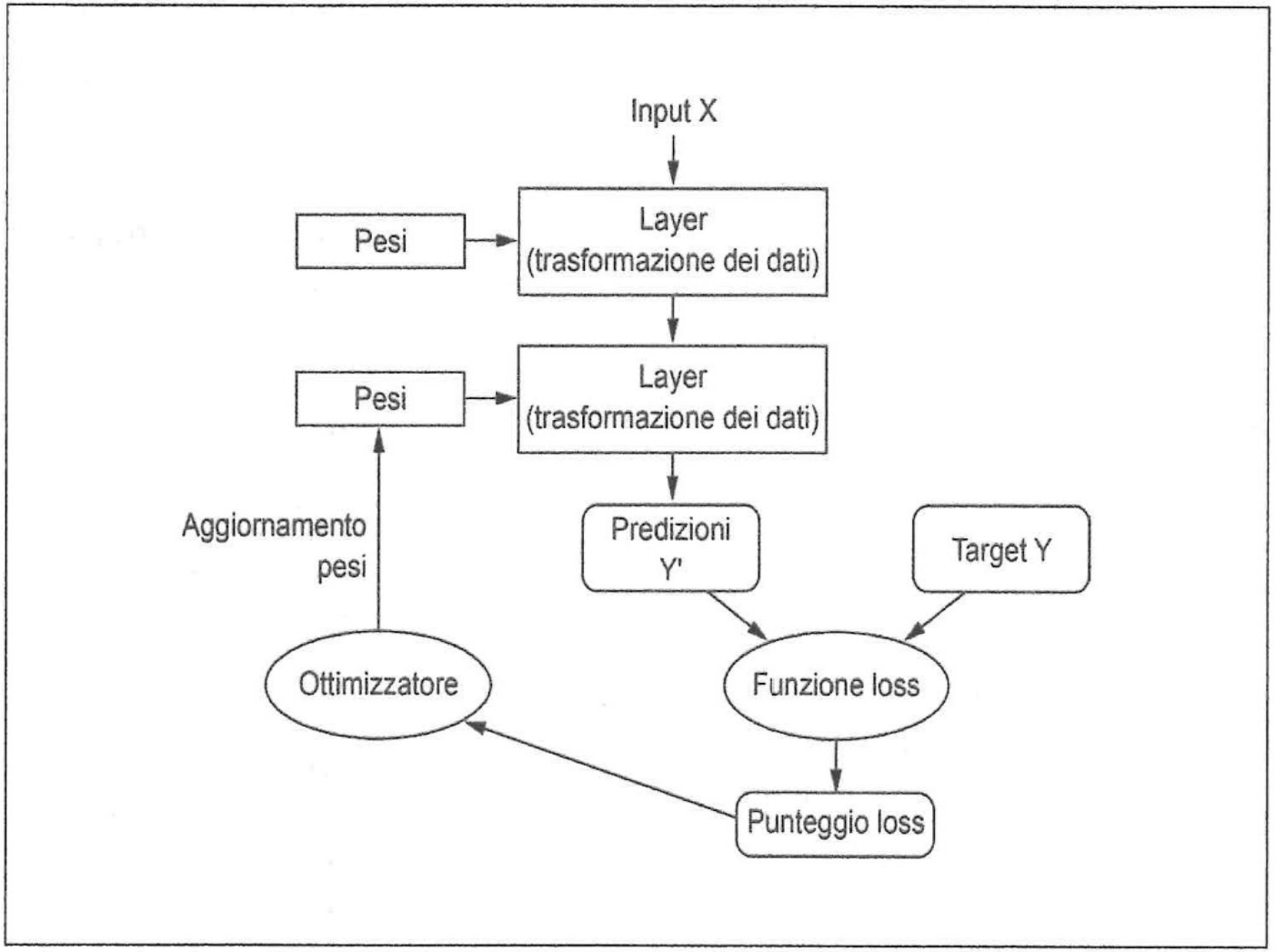


Figura 3.1 Relazioni fra la rete, i layer, la funzione obiettivo e l'ottimizzatore.

Anatomia di una rete neurale

- Si misura quanto l'output si avvicina al risultato previsto attraverso la funzione loss che calcola un punteggio di distanza, che valuta le prestazioni della rete.

- I valori dei pesi vengono aggiustati al fine di ridurre la distanza attraverso l'ottimizzatore.

- Ai pesi iniziali vengono assegnati valori casuali che via via vengono aggiustati sempre più verso la direzione corretta, riducendo così la distanza fra le predizioni e il target.

- Questo procedimento prende il nome di *ciclo di addestramento* che ripetuto un numero sufficiente di volte fornisce valori di pesi che minimizzano la funzione loss.

Lavorare con dati testuali

- Ambiti di applicazione:
  - Classificazione recensioni di film come positive o negative
  - Identificazione dell'autore di un libro
  - Rispondere a determinate domande
  - Estrarre informazioni rilevanti

- I modelli di deep learning non prendono come input un testo grezzo ma operano su valori numerici.

- Il processo di trasformazione del testo in valori numerici è detto vettorizzazione.

Trasformazione dei dati
Bag-of-words

- E’ una rappresentazione del testo che descrive l'occorrenza di parole all'interno di un documento,
- Tende ad essere impiegato nei modelli «superficiali» di elaborazione del linguaggio non nei modelli di deep learning

ISPRO
Istituto per lo studio, la prevenzione e la rete oncologica

Trasformazione dei dati
Word vectors

- Detto anche Word Embedding, consiste nel trasformare il testo in vettori numerici.

- Questa trasformazione è necessaria perché molti algoritmi di apprendimento automatico (comprese le reti profonde) richiedono che i loro input siano vettori di valori continui; semplicemente non funzionano su stringhe di testo normale.

- Oltre ad essere suscettibile di elaborazione mediante algoritmi di apprendimento, questa rappresentazione vettoriale ha due proprietà importanti e vantaggiose:
  - Riduzione della dimensionalità
  - Somiglianza contestuale

Trasformazione dei dati
Word vectors

- I word vectors vengono utilizzati per l'analisi semantica, al fine di estrarre il significato dal testo per consentire la comprensione del linguaggio naturale.

- Affinché un modello linguistico sia in grado di prevedere il significato del testo, deve poter riconoscere la somiglianza contestuale delle parole.

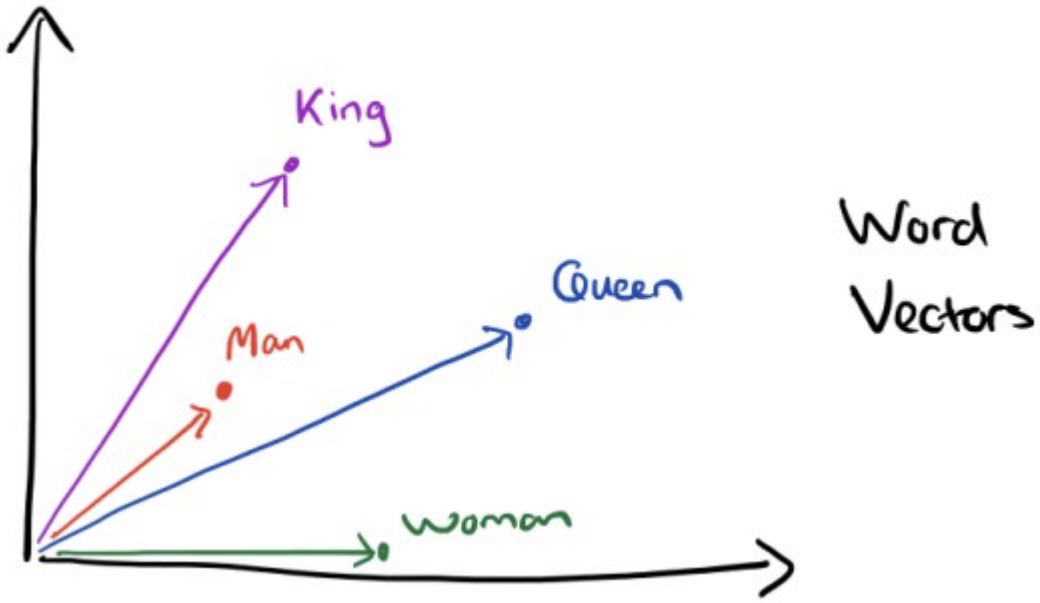
- Se ad esempio, l'obiettivo fosse riconoscere un termine riferito ad un *frutto* (come *mela o arancia*) nelle frasi in cui compaiono parole che fanno riferimento a *coltivare, raccogliere, mangiare o spremere*, sarebbero da considerarsi sicuramente più informative piuttosto che ad esempio la parola *aereo*.

- L'algoritmo richiede in ingresso un testo e restituisce un insieme di vettori che rappresentano la distribuzione semantica delle parole nel testo. Per ogni parola viene costruito un vettore in modo da rappresentarla come un punto nello spazio. In questo spazio le parole saranno più vicine se riconosciute come semanticamente più simili.

Trasformazione dei dati
Word vectors

- Eseguendo semplici operazioni algebriche sui vettori delle parole, è stato mostrato ad esempio che

  - vettore ("Re") - vettore ("Uomo") + vettore ("Donna") ottiene come risultato un vettore più vicino alla rappresentazione vettoriale della parola ("Regina") .



Apprendimento supervisionato

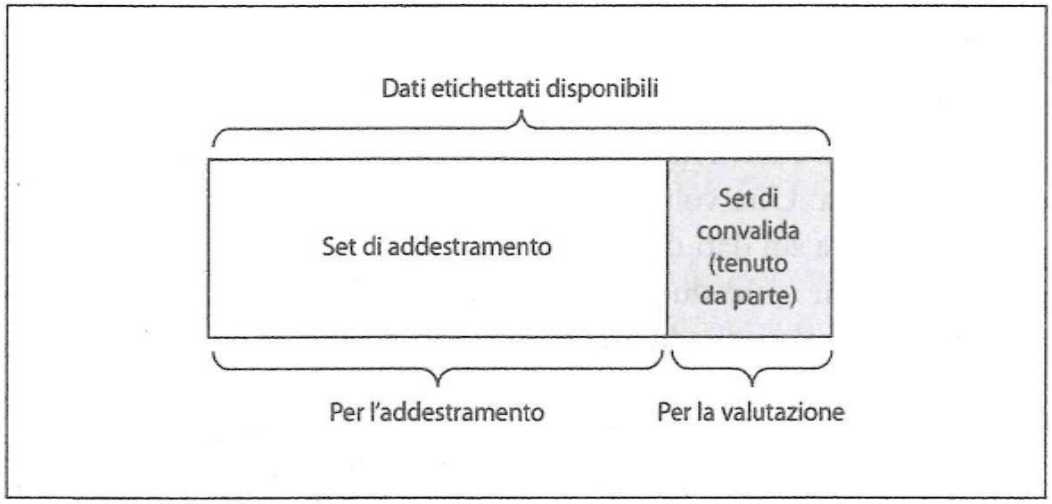


Figura 4.1 Semplice suddivisione per la convalida hold-out.

- L'obiettivo è quello di imparare la relazione fra gli input di addestramento e i target di addestramento
- Consiste nell'imparare a mappare i dati di input su determinati target noti (chiamati anche annotazioni), in base a un insieme di esempi (annotati da esseri umani)

Esperienze e risultati preliminari

ISPRO
Istituto per lo studio, la prevenzione e la rete oncologica

IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, ACCEPTED JUNE 2020 1

Classification of cancer pathology reports: a large-scale comparative study

Stefano Martina, Leonardo Ventura, and Paolo Frasconi

Abstract—We report about the application of state-of the-art deep learning techniques to the automatic and interpretable assignment of ICD-O3 topography and morphology codes to free-text cancer reports. We present results on a large dataset (more than 80 000 labeled and 1 500 000 unlabeled anonymized reports written in Italian and collected from hospitals in Tuscany over more than a decade) and with a large number of classes (134 morphological classes and 61 topographical classes). We compare alternative architectures in terms of prediction accuracy and interpretability and show that our best model achieves a multiclass accuracy of 90.3% on topography site assignment and 84.8% on morphology type assignment. We found that in this context hierarchical models are not better than flat models and that an element-wise maximum aggregator is slightly better than attentive models on site classification.

IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, ACCEPTED JUNE 2020 1

Classification of cancer pathology reports: a large-scale comparative study

Stefano Martina, Leonardo Ventura, and Paolo Frasconi

Topografia

C _ _ . _

- Prime due cifre descrivono la sede
- La terza cifra descrive la sotto-sede

Es. *C50.2* quadrante supero-interno(2) della mammella(50)

Morfologia

_ _ _ _ / _

- Prime quattro cifre descrivono l'istologia
- La quinta cifra descrive il comportamento

Es. *8140/3* è un adenocarcinoma (adeno 8140; carcinoma 3)

ISPRO
Istituto per lo studio, la prevenzione e la rete oncologica

Dati

- Sono stati analizzati circa 1.500.000 di referti anatomo-patologici dal Registro Toscano Tumori per il periodo 1990-2014.
- Di questi circa il 6% aveva un referto di diagnosi di tumore con le codifiche dei codici morfologici e topografici identificati dal personale del Registro Tumori.
- La restante parte dei referti era riferita a tessuti non cancerosi.
- I referti AP contenevano informazioni su macroscopia, diagnosi e in alcuni casi anamnesi familiare.
- 61 classi topografiche e 134 classi morfologiche

Risultati

- Predizione del codice topografico (61 classi)
- Predizione del codice morfologico (134 classi)

	Topography				Morphology			
	Acc.	Top 3	Top 5	MacroF1	Acc.	Top 3	Top 5	Macro F1
U-SVM	89.7	95.9	96.8	60.0	82.4	94.0	95.6	53.7
B-XGB	89.1	95.8	97.2	58.0	84.1	94.4	96.5	59.6
G-GRU	89.9	96.5	97.7	58.3	83.3	94.6	96.6	55.2
BERT	89.9	96.3	97.8	56.6	84.3	93.2	94.9	51.1
G-MAXi	88.0	95.4	96.2	46.1	73.4	91.0	93.6	31.3
G-MAXh	89.9	96.2	97.8	58.8	83.7	94.4	96.4	54.5
G-ATTh	89.9	96.3	97.7	58.0	83.7	94.4	96.2	57.5
G-MAX	**90.3**	**96.6**	**98.1**	**61.9**	84.6	**95.0**	**96.9**	59.2
G-ATT	90.1	96.2	97.6	60.0	**84.8**	94.9	**96.9**	**61.3**

Risultati

- Easy: classi con più di 1000 esempi nel test set (colon-retto, polmone, mammella, prostata)
- Average: classi che avevano fra 100 e 1000 esempi nel dataset
- Hard: classi con meno di 100 esempi nel dataset

	Topography			Morphology		
	easy ($1000 < s$) (4 cls)	avg. ($100 < s \leq 1000$) (18 cls)	hard ($s \leq 100$) (39 cls)	easy ($1000 < s$) (5 cls)	avg. ($100 < s \leq 1000$) (18 cls)	hard ($s \leq 100$) (111 cls)
U-SVM	95.7	**86.9**	50.9	90.5	68.6	48.4
B-XGB	95.6	86.4	48.2	92.0	72.4	54.8
G-GRU	**96.1**	72.2	48.0	91.4	71.6	49.7
BERT	95.7	73.2	44.9	**92.9**	**74.4**	43.9
G-MAXi	95.0	66.6	31.4	87.1	41.9	25.1
G-MAXh	95.8	72.4	48.8	92.7	71.8	48.8
G-ATTh	96.0	73.1	47.1	91.9	72.3	52.6
G-MAX	96.0	73.3	**53.1**	92.7	72.3	53.8
G-ATT	96.0	73.1	50.3	92.8	72.3	**56.7**

ISPRO
Istituto per lo studio, la prevenzione e la rete oncologica

Conclusioni

- L’intelligenza artificiale si sta ormai sempre più diffondendo in ambito medico in generale ed oncologico in particolare.
- Rappresenta uno strumento che ben si adatta a contesti con grandi moli di dati con risultati molto promettenti.
- E’ naturale immaginare un futuro in cui l’intelligenza artificiale entrerà con sempre maggior forza come strumento di supporto tanto per il clinico quanto per il sistema sanitario in genere.
- Necessità di utilizzo di tecniche automatiche per velocizzare la lettura dei referti ed estrarre variabili non codificate dal Registro Tumori.