



UNIVERSITÀ DEGLI STUDI DI NAPOLI  
**FEDERICO II**

# Cognitive Computing System applicato allo studio del Glioblastoma

*Docenti:*

Prof. Paolo Maresca

*Autori:*

Francesca Lettieri - M63000956

Angelo Russo - M63001016

Guido Maria Secondulfo - M63000927

# Indice

<b>1</b>	<b>Introduzione</b>	<b>3</b>
1.1	Traccia . . . . .	3
<b>2</b>	<b>Architettura</b>	<b>3</b>
2.1	Tecnologie utilizzate . . . . .	3
2.1.1	E-utilities . . . . .	3
2.1.2	IBM Cloud Db2 . . . . .	3
2.1.3	GitHub . . . . .	3
2.1.4	IBM Cloud Foundry . . . . .	3
2.1.5	IBM Watson Machine Learning . . . . .	4
2.1.6	IBM Watson Knowledge Studio . . . . .	4
2.1.7	fastai . . . . .	4
2.1.8	IBM Cloud API . . . . .	4
2.1.9	Jupyter . . . . .	4
<b>3</b>	<b>Implementazione</b>	<b>5</b>
3.1	Flusso di esecuzione del sistema globale . . . . .	5
<b>4</b>	<b>Apprendimento sugli articoli medici tramite NLP</b>	<b>6</b>
4.1	Flusso di esecuzione del sistema di retrieval degli articoli . . . . .	6
4.1.1	E-utilities . . . . .	6
4.1.2	Query di ricerca . . . . .	7
4.1.3	GitHub e IBM Cloud Foundry . . . . .	9
4.1.4	IBM Cloud Db2 . . . . .	9
4.1.5	Funzioni implementate . . . . .	14
4.2	Flusso di esecuzione del sistema di apprendimento sugli articoli .	16
4.3	IBM Watson Knowledge Studio . . . . .	16
4.3.1	Import degli Articoli . . . . .	16
4.3.2	Creazione del Modello Entità-Relazione . . . . .	16
4.4	Valutazione del modello . . . . .	23
<b>5</b>	<b>Apprendimento sulle Risonanze Magnetiche tramite Visual Recognition</b>	<b>27</b>
5.1	Flusso di esecuzione del sistema di apprendimento sulle MRI . .	27
5.2	Visual Recognition di Watson Studio . . . . .	27
5.2.1	Implementazione . . . . .	28
5.2.2	Risultati ottenuti - Primo Ciclo . . . . .	31
5.2.3	Risultati ottenuti - Secondo Ciclo . . . . .	33
5.2.4	Risultati ottenuti - Terzo Ciclo . . . . .	35
5.2.5	Risultati ottenuti - Quarto Ciclo . . . . .	37
<b>6</b>	<b>Considerazioni e Sviluppi Futuri</b>	<b>39</b>
<b>A</b>	<b>RadiAnt</b>	<b>41</b>
A.1	Pre-processing effettuato tramite RadiAnt . . . . .	42

<b>B</b>	<b>Evaluation Metrics</b>	<b>44</b>
B.1	Accuracy . . . . .	44
B.2	Precision e Recall . . . . .	44
B.3	F1-Score . . . . .	45
<b>C</b>	<b>fastai</b>	<b>46</b>
C.1	CNN con la libreria <i>fastai</i> . . . . .	46
	C.1.1 Implementazione . . . . .	46
	C.1.2 Risultati ottenuti . . . . .	48
<b>D</b>	<b>Rimozione testi pediatrici</b>	<b>51</b>

# 1 Introduzione

## 1.1 Traccia

Il progetto da noi sviluppato si articola in due percorsi distinti:

1. Retrieval da PubMed di articoli inerenti al Glioblastoma Multiforme (GBM), creazione di un database atto a contenere le informazioni rilevanti ed, infine, creazione di un modello di NLP (Natural Language Processing) al fine di realizzare un sistema di apprendimento in grado di realizzare una standard meta-analysis accurata;
2. Acquisizione delle risonanze magnetiche (MRI) di pazienti affetti da GBM e costruzione di un modello di apprendimento su di esse.

# 2 Architettura

## 2.1 Tecnologie utilizzate

### 2.1.1 E-utilities

Le E-utilities (Entrez Programming Utilities) costituiscono un insieme di 8 programmi server-side che forniscono un'interfaccia stabile per il sistema di database e query Entrez. Le E-utilities utilizzano una sintassi a URL fisso che traduce un insieme di parametri di input nei valori necessari ai componenti software del NCBI per ricercare e recuperare i dati richiesti.

### 2.1.2 IBM Cloud Db2

IBM Db2 on Cloud è un database cloud transazionale progettato per offrire prestazioni elevate. Esso assicura la scalabilità orizzontale e verticale per soddisfare le richieste di business.

### 2.1.3 GitHub

Git è un software di controllo versione distribuito utilizzabile da interfaccia a riga di comando, sviluppato per gestire anche progetti di elevate dimensioni con elevate velocità ed efficienza.

### 2.1.4 IBM Cloud Foundry

Cloud Foundry è una PaaS (platform-as-a-service) open source su IBM Cloud che consente di implementare e scalare le app senza gestire i server. È stato da noi utilizzato al fine di poter eseguire uno script in linguaggio Python all'interno dell'ambiente IBM Cloud.

### **2.1.5 IBM Watson Machine Learning**

IBM Watson Machine Learning aiuta i data scientist e gli sviluppatori ad accelerare l'implementazione di AI e machine-learning. Con i suoi modelli estensibili, aperti, Watson Machine Learning aiuta le aziende a semplificare e sfruttare l'AI in scala su qualsiasi cloud.

### **2.1.6 IBM Watson Knowledge Studio**

IBM Watson consente l'apprendimento del linguaggio del dominio con modelli personalizzati che identificano le entità e le relazioni specifiche del settore in testo non strutturato. I modelli possono essere creati in un ambiente collaborativo progettato sia per gli sviluppatori che per gli esperti di dominio, senza il bisogno di scrivere codice, e possono essere utilizzati in IBM Watson Discovery, IBM Watson Natural Language Understanding e IBM Watson Explorer.

### **2.1.7 fastai**

fastai è una libreria di deep learning che fornisce agli operatori componenti di alto livello in grado di fornire rapidamente e facilmente risultati di ultima generazione in domini di deep learning standard, e fornisce ai ricercatori con componenti di basso livello che possono essere mescolati e abbinati per costruire nuovi approcci. fastai mira a fare entrambe le cose senza compromessi sostanziali in facilità d'uso, flessibilità, o prestazioni. Questo è possibile grazie ad un'architettura a strati, che esprime schemi comuni di base di molte tecniche di deep learning e di elaborazione dati in termini di astrazioni disaccoppiate. Queste astrazioni possono essere espresse in modo conciso e chiaro sfruttando il dinamismo del linguaggio Python sottostante e la flessibilità della libreria Pytorch.

### **2.1.8 IBM Cloud API**

Le IBM Cloud API sono delle API che consentono di garantire l'accesso ai servizi IBM Cloud ai propri servizi ed alle proprie applicazioni che lo necessitano.

### **2.1.9 Jupyter**

Il Progetto Jupyter è un'organizzazione creata per sviluppare software open-source, e supportare ambienti di esecuzione in decine di lingue.

JupyterLab offre un'interfaccia per la gestione di documenti di diverso formato, per attività di text editing e attività da terminale.

### 3 Implementazione

#### 3.1 Flusso di esecuzione del sistema globale

Il diagramma di flusso rappresentante l'elaborazione effettuata è il seguente:

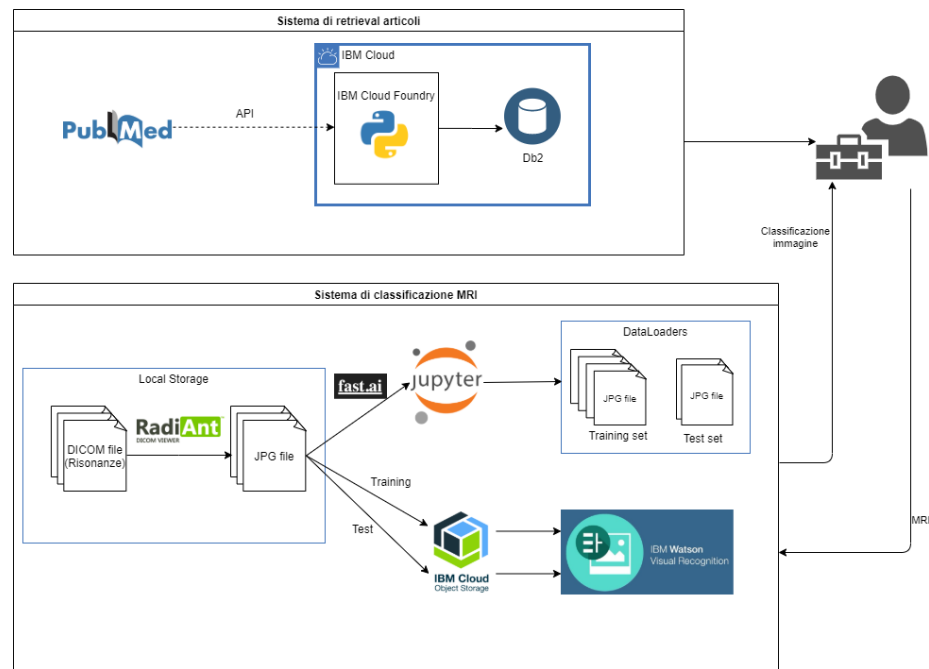


Figura 1: Diagramma di flusso dell'esecuzione

Al fine di agevolare la trattazione, affronteremo i problemi separatamente.

## 4 Apprendimento sugli articoli medici tramite NLP

### 4.1 Flusso di esecuzione del sistema di retrieval degli articoli

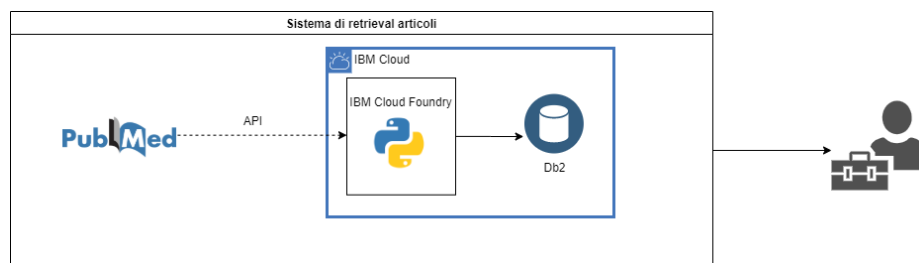


Figura 2: Diagramma di flusso dell'esecuzione

Innanzitutto è stato effettuato il download degli articoli desiderati da PubMed. Dopo una prima elaborazione di tali documenti, essi sono stati memorizzati in un database relazionale all'interno dell'ambiente IBM Cloud.

#### 4.1.1 E-utilities

In seguito ad un confronto con gli esperti di dominio, abbiamo ritenuto opportuno effettuare un retrieval di articoli provenienti unicamente da PubMed, in quanto source affidabile e attendibile.

A tal fine è stato sviluppato uno script in linguaggio Python (*pubmed.py*) che, sfruttando le E-utilities, è in grado di effettuare il retrieval degli articoli desiderati da PubMed stesso.

In particolare, la funzione *getPubMedIDArticles* recupera gli ID degli articoli desiderati da PubMed, in base ai termini specificati. L'implementazione di tale funzione è la seguente:

```
def getPubMedIDArticles(query):
    test = "esearch.fcgi?db=pubmed"
    term = "&term="+query #query della ricerca
    maxArts = "&retmax=100000" #Numero massimo di articoli
    recuperati dalla ricerca
    link = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/
    /%22+test+term+maxArts

    s = urllib.request.urlopen(link)
    contents = s.read()
    file = open("export_idArticles.xml", 'wb')
    file.write(contents)
    file.close()
```

#### 4.1.2 Query di ricerca

La query effettuata è stata elaborata dagli esperti di dominio. Essa espone quali siano i termini da ricercare all'interno di tutti i campi di un articolo, inclusi i MeSH Terms <sup>1</sup> associati all'articolo stesso.

I criteri di inclusione per la selezione degli articoli sono i seguenti:

- Studi randomizzati di confronto fra trattamenti;
- Fase II o III (compresi studi Fase I/II purchè contengano risultati adeguati della Fase II);
- OS sopravvivenza globale (overall survival) e PFS sopravvivenza libera da progressione (progression free survival) e ORR (Overall Response Rate) Tasso di risposta complessiva;
- Solo Glioblastoma multiforme (GBM) o High Grade Glioma in ripresa dopo pregressa chirurgia e terapia adiuvante o in progressione.

I criteri di esclusione per la selezione degli articoli sono, invece, i seguenti:

- Tumori pediatrici;
- Studi non randomizzati;
- Studi a singolo braccio;
- Studi di fase I;
- GBM in trattamento adiuvante;
- Review/metanalisi;
- Studi preclinici;
- Studi di farmacoeconomia;
- Studi di confronto senza analisi statistiche.

La formulazione risultante della query è, dunque, la seguente:

---

<sup>1</sup>Medical Subject Headings (acronimo: MeSH) è un enorme vocabolario controllato (o sistema di metadati) ideato con l'obiettivo di indicizzare la letteratura scientifica in ambito biomedico. Il tesoro è stato creato dalla National Library of Medicine (NLM) degli Stati Uniti, che è responsabile anche della sua gestione. Il MeSH viene adoperato per l'indicizzazione degli articoli delle oltre 5000 riviste mediche presenti nel database bibliografico Medline/PubMed e nel catalogo dei libri della NLM. Il vocabolario può essere consultato e scaricato gratuitamente da tutti gli utenti di internet. La stampa dell'edizione cartacea è stata abbandonata nel 2007.



(glioblastoma OR glioblastoma multiforme OR anaplastic astocytoma OR malignant glioma OR high grade glioma OR high grade astrocytoma) AND (diffuse OR multiple OR unresectable OR advanced OR recurrent OR inoperable OR disseminated OR progressive OR relapsed) AND (clinical outcome OR OS OR PFS OR ORR) AND (chemotherapy OR target Therapy OR immunotherapy OR cyclophosphamide OR bevacizumab OR temozolomide OR irinotecan OR carmustine OR lomustine OR nitrosourea OR BCNU OR CCNU OR carboplatin OR cisplatin OR platinum OR PCV OR etoposide OR VP-16 OR cilengitide OR alternating electric field therapy OR TTF OR tumor treating fields OR procarbazine OR vincristine OR sunitinib OR lapatinib OR pazopanib OR temsirolimus OR everolimus OR fotemustine OR sagopilone OR Bortezomib OR Vorinostat OR enzastaurin OR Vandetanib OR veliparib OR dasatinib OR ultrafractionated RT OR nivolumab OR Regorafenib OR pembrolizumab OR ipilimumab OR radiosurgery OR ypo-fractioned radiation OR anlotinib OR Cesium-131 brachytherapy OR gefitinib OR sorafenib OR dacomitinib OR pidilizumab OR ivosidenib OR Interstitial brachytherapy OR erlotinib OR depatuxizumab mafodotin OR durvalumab OR onartuzumab OR rituximab OR Disulfiram OR reirradiation OR capmatinib OR buparlisib OR drug delivery tecnology OR convection-enhanced delivery OR bulk flow OR intra-arterial chemotherapy OR nanotechnology-based controlled delivery OR intraoperative polymer implants OR dagnetic cationic microsphere OR biodegradable polymer wafers OR lipid-coated microbubbles OR magnetic hyperthermia OR photodynamic therapy)

Figura 3: Query originale

Tale query è stata, quindi, da noi raffinata al fine di poter effettuare il retrieval degli articoli che soddisfacevano i criteri richiesti.

L'operazione di raffinamento è stata dettata dalla risposta delle API di PubMed a determinati termini (es. "ypo-fractioned", "dagnetic") che facevano parte della query iniziale; tali termini, omessi in fase di raffinamento, non risultavano presenti all'interno del tesoro (i.e. il MeSH).

(glioblastoma+OR+glioblastoma+multiforme+OR+anaplastic+astocytoma+OR+malignant+glioma+OR+high+grade+glioma+OR+high+grade+astrocytoma)+AND+(diffuse+OR+multiple+OR+unresectable+OR+advanced+OR+recurrent+OR+inoperable+OR+disseminated+OR+progressive+OR+relapsed)+AND+(clinical+outcome+OR+OS+OR+PFS+OR+ORR)+AND+(chemotherapy+OR+target+Therapy+OR+immunotherapy+OR+cyclophosphamide+OR+bevacizumab+OR+temozolomide+OR+irinotecan+OR+carmustine+OR+lomustine+OR+nitrosourea+OR+BCNU+OR+CCNU+OR+carboplatin+OR+cisplatin+OR+platinum+OR+PCV+OR+etoposide+OR+VP-16+OR+cilengitide+OR+alternating+electric+field+therapy+OR+TTF+OR+tumor+treating+fields+OR+procarbazine+OR+vincristine+OR+sunitinib+OR+lapatinib+OR+pazopanib+OR+temsirolimus+OR+everolimus+OR+fotemustine+OR+sagopilone+OR+Bortezomib+OR+Vorinostat+OR+enzastaurin+OR+Vandetanib+OR+veliparib+OR+dasatinib+OR+ultrafractionated+RT+OR+nivolumab+OR+Regorafenib+OR+pembrolizumab+OR+ipilimumab+OR+radiosurgery+OR+ypo-fractioned+radiation+OR+anlotinib+OR+Cesium-131+brachytherapy+OR+gefitinib+OR+sorafenib+OR+dacomitinib+OR+pidilizumab+OR+ivosidenib+OR+Interstitial+brachytherapy+OR+erlotinib+OR+depatuxizumab+mafodotin+OR+durvalumab+OR+onartuzumab+OR+rituximab+OR+Disulfiram+OR+reirradiation+OR+capmatinib+OR+buparlisib+OR+drug+delivery+tecnology+OR+convection-enhanced+delivery+OR++bulk+flow+OR+intra-arterial+chemotherapy+OR+nanotechnology-based+controlled+delivery+OR+intraoperative+polymer+implants+OR+dagnetic+cationic+microsphere++OR+biodegradable+polymer+wafers+OR+lipid-coated+microbubbles+OR+magnetic+hyperthermia+OR+photodynamic+therapy)

Figura 4: Query Raffinata

In entrambi i casi, il numero di articoli recuperati è pari a 1'318, per questo l'eliminazione dei termini è risultata ininfluente per quanto riguarda il risultato finale ottenuto. Un ulteriore filtraggio degli articoli recuperati è stato fatto in base alla data di pubblicazione, secondo quanto richiesto e consigliato dagli esperti di dominio: escludendo gli articoli pubblicati prima del 1990, il numero finale di articoli recuperati è pari a 1'313.

#### 4.1.3 GitHub e IBM Cloud Foundry

Per poter importare gli articoli in ambiente IBM Cloud, innanzitutto è stato creato un repository su GitHub all'interno del quale è stato memorizzato il file *pubmed.py*.

In particolare, sono state effettuate le operazioni di **add**, **commit** e **push** del file all'interno del repository attraverso il prompt dei comandi.

#### 4.1.4 IBM Cloud Db2

Innanzitutto è stata istanziata una versione lite del Database Db2. Utilizzando le credenziali di accesso fornite dal sistema e le API IBM\_DB, è stato possibile accedere al Database attraverso una funzione all'interno del file *pubmed.py*.

Per la modellazione del database, volendo tener traccia di un'unica entità, gli articoli, è stato realizzato il seguente diagramma ER<sup>2</sup>:

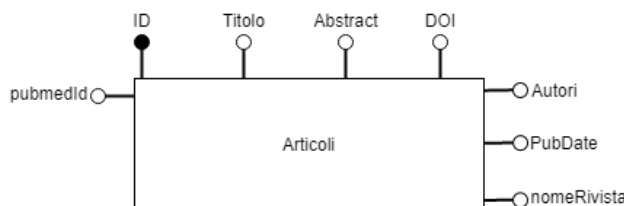


Figura 5: Diagramma ER

Abbiamo ipotizzato di memorizzare, per ogni articolo:

- DOI (Digital Object Identifier) - quando disponibile;
- ID dell'articolo nel database PubMed;
- Titolo;
- Lista di autori - quando disponibile;

<sup>2</sup>In informatica, nell'ambito della progettazione dei database, il modello entity-relationship (detto anche modello E-R; in italiano: modello entità-associazione o modello entità-relazione) è un modello teorico per la rappresentazione concettuale e grafica dei dati a un alto livello di astrazione, formalizzato dal prof. Peter Chen nel 1976.

Il modello entità-associazione viene spesso utilizzato nella prima fase della progettazione di una base di dati, nella quale è necessario tradurre le informazioni risultanti dall'analisi di un determinato dominio in uno schema concettuale, chiamato schema E-R (schema entità-associazione) o diagramma E-R (diagramma entità-associazione).

- Data di pubblicazione - quando disponibile;
- Rivista di pubblicazione - quando disponibile;
- Abstract - quando disponibile.

Osserviamo che, durante l'attività di recupero dalle API Pubmed, il DOI era assente in alcuni articoli, il che lo rendeva inadeguato al ruolo di chiave primaria dell'entità. Per questo motivo, abbiamo utilizzato un ulteriore identificativo numerico con auto-incremento come chiave primaria.

Quindi, è stato possibile creare una tabella ARTICOLIGBM attraverso la seguente query:

```
CREATE TABLE articoligbm (
  id INTEGER NOT NULL PRIMARY KEY GENERATED ALWAYS AS
    IDENTITY (START WITH 1 INCREMENT BY 1) ,
  pubmedId INTEGER UNIQUE NOT NULL,
  doi VARCHAR(80) ,
  titolo VARCHAR(1000) ,
  autori VARCHAR(5000) ,
  dataPubblicazione DATE,
  abstract VARCHAR(10000) ,
  nomeRivista VARCHAR (1000)"
)
```

La connessione al database è realizzata attraverso la funzione *connectionDB2*, mentre tale connessione viene chiusa attraverso la funzione *endConnectionDB2*. L'implementazione di entrambe le funzioni è riportata di seguito:

```
def connectionDB2():
  conn_info = ""DATABASE=BLUDB;HOSTNAME=dashdb-txn-sbox
    -yp-lon02-13.
    services.eu-gb.bluemix.net;PORT=50000;PROTOCOL=TCPIP;
    UID=xxd26106;PWD=s5bzk6p^3s94hght;"
  conn = ibm_db.connect(conn_info , "", "")
  return conn
```

```
def endConnectionDB2(conn):
  ibm_db.close(conn)
```

Si noti che le informazioni relative al Database, quali nome dello stesso, l'Hostname, il porto, il protocollo, l'username e la password, sono strettamente legate all'account IBM Cloud che ha istanziato il database. In caso di migrazione del database, o di creazione di uno nuovo, è necessario modificare l'intera variabile *conn\_info* con le rispettive Credenziali di Servizio generate sul portale IBM Cloud.

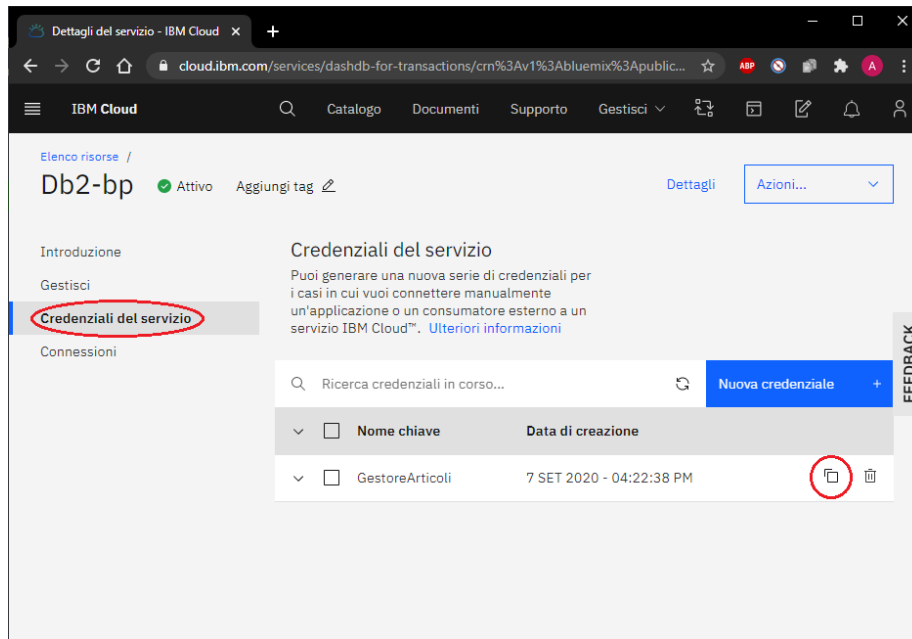


Figura 6: Schermata DB2 per ottenere le credenziali del servizio

Infine, il popolamento del database può essere effettuato attraverso due diverse funzioni. La prima, *insertArticleDB2*, inserisce un singolo articolo all'interno del database attraverso la seguente query sql:

```
INSERT INTO articoli (pubmedId, doi, titolo , autori ,
                    dataPubblicazione, abstract, nomeRivista)
```

```
VALUES (valore_pubmedId , valore_doi , valore_titolo ,
        valore_autori ,
        valore_dataPubblicazione , valore_abstract ,
        valore_nomeRivista);
```

La seconda, *insertArrayDB2*, è quella che è stata utilizzata per realizzare il primo popolamento del database. Essa riceve in ingresso un array di articoli, andando ad inserire ognuno di essi nel database stesso attraverso la stessa query vista precedentemente.

In entrambi i casi, si è resa, quindi, necessaria la definizione di una classe Articolo, la cui implementazione è riportata di seguito:

```
class Articolo():
    def __init__(self, titolo , autori , pubDate, pubMedId,
                doi, abstract, journalTitle):
        self.titolo = titolo
        self.autori = autori
        self.abstract = abstract
```

```

        self.doi = doi
        self.pubMedId = pubMedId
        self.pubDate = pubDate
        self.journalTitle = journalTitle
    def __str__(self):
        return "PUBMED ID: %s\nDOI: %s\nTitolo: %s\
            nAutori: %s\nData di Pubblicazione: %s\
            nAbstract: %s\nJournal Title: %s\n\n" % (self.
            pubMedId, self.doi, self.titolo, ', '.join(
            self.autori), self.pubDate, self.abstract,
            self.journalTitle)

```

L'implementazione delle funzioni sopra menzionate è la seguente:

```

def insertArticleDB2 (articolo):
    abstract_temp = " "
    titolo_temp = " "
    jtitle_temp = " "
    autori_insert = ', '.join(articolo.autori)
    if (autori_insert != None):
        autori_insert = autori_insert.replace(" ", " ")

    if (articolo.titolo != None):
        titolo_temp = articolo.titolo.replace(" ", " ")

    if (articolo.abstract != None):
        abstract_temp = articolo.abstract.replace(" ", " ")

    if (articolo.journalTitle != None):
        jtitle_temp = articolo.journalTitle.replace(" ", " ")

    insert = ibm_db.exec_immediate(conn, "INSERT INTO
    articoli (pubmedId, doi, titolo, autori,
    dataPubblicazione, abstract, nomeRivista) VALUES
    ('"+articolo.doi+"', '"+titolo_temp+"', '"+
    autori_insert+"', '"+ articolo.pubDate+"', '"+
    abstract_temp+"', '"+jtitle_temp+"')")

    print("[ "+str(i)+" ] - Inserito\n") #DEBUG dell'
    articolo inserito

```

```

def insertArrayDB2 (articoli):
    for articolo in articoli:
        abstract_temp = " "
        titolo_temp = " "

```

```

jtitle_temp = " "

autori_insert = ', '.join(articolo.autori)
if (autori_insert != None):
    autori_insert = autori_insert.replace(" ", "
    ")

if (articolo.titolo != None):
    titolo_temp = articolo.titolo.replace(" ", "
    ")

if (articolo.abstract != None):
    abstract_temp = articolo.abstract.replace(" ", "
    ")

if (articolo.journalTitle != None):
    jtitle_temp = articolo.journalTitle.replace("
    ", " ")

insert = ibm_db.exec_immediate(conn, "INSERT INTO
    articoli (pubmedId, doi, titolo, autori,
    dataPubblicazione, abstract, nomeRivista)
    VALUES ('"+articolo.doi+"', '"+titolo_temp+"',
    '"+ autori_insert+"', '"+ articolo.pubDate+"
    ', '"+abstract_temp+"', '"+jtitle_temp+" ')")

print("[ "+str(i)+" ] - Inserito\n") #DEBUG dell'
    articolo inserito

```

Inizialmente, si era pensato di mantenere in memoria le informazioni degli articoli recuperati, per poi caricarle tutte insieme sul database: questo metodo però è risultato poco efficiente a causa della grande mole di articoli da caricare. Tale risultato è stato ottenuto stressando il sistema tramite una query generica di prova, che cercasse gli articoli che contenessero in qualsiasi campo (sia esso Titolo, MeSH o Abstract) il termine "glioblastoma". Questa query di prova ha prodotto come risultato ben 41'364 articoli e pubblicazioni, rallentando di molto il processo di recupero, il mantenimento temporaneo in memoria e infine l'inserimento nel Database degli stessi. È risultato molto più conveniente, piuttosto, gestire grosse moli di articoli recuperandoli e inserendoli immediatamente nel database, in modo da non perdere l'eventuale cluster di articoli in caso di disconnessioni causate dal server di PubMed (problema che, più volte, è stato riscontrato durante il testing dello script). Conviene utilizzare, quindi, la funzione *insertArrayDB2* quando vengono recuperati quantità minime di articoli (nell'ordine massimo delle centinaia), come nel caso della query utilizzata ed esposta nel sottoparagrafo 3.1.2, dove la query ha restituito 1'318 articoli come esito della ricerca.

#### 4.1.5 Funzioni implementate

Dopo aver ottenuto la lista contenente gli ID degli articoli, la funzione *pubMedIDConverter* recupera le informazioni ad essi associate. L'implementazione di tale funzione è la seguente:

```
def PubMedIdConverter(articoli):
    data_file = 'export.xml'
    tree = ET.parse(data_file)
    root = tree.getroot()
    id_converter = "https://eutils.ncbi.nlm.nih.gov/
        entrez/eutils/efetch.fcgi?db=pubmed&id="
    i=0

    for child in root.findall("./IdList/Id"):
        i=i+1
        id = child.text
        id_converter1 = id_converter+id+"&retmode=xml"

        xml = urllib.request.urlopen(id_converter1).read()
        xml_tree = ET.fromstring(xml)    # Root

        for child2 in xml_tree.findall("PubmedArticle"):
            temp_title = ""
            temp_abstract = ""
            temp_author = ""
            temp_doi = "Mancante"
            temp_pubdate = ""
            autori_vector = []

            for child2_0 in child2.findall("
                MedlineCitation/Article"):
                for child2_JournalTitle in child2_0.
                    findall("Journal/Title"):
                        temp_jtitle = child2_JournalTitle.
                            text
                for child2_1 in child2_0.findall("
                    ArticleTitle"):
                        temp_title = child2_1.text
                for child2_2 in child2_0.findall("
                    Abstract/AbstractText"):
                        temp_abstract = child2_2.text
                for child2_3 in child2_0.findall("
                    AuthorList/Author"):
                        #La presente soluzione e' concepita
                            nel caso si voglia tenere traccia
```

```

        solo di iniziali del nome e
        cognome (per intero)
path_cognome = child2_3.find("
    LastName")
path_iniziali = child2_3.find("
    Initials")
if (path_cognome != None):
    if (path_iniziali != None):
        temp_author = child2_3.find("
            LastName").text + " " +
            path_iniziali.text + "."
    else:
        temp_author = child2_3.find("
            LastName").text
    autori_vector.append(temp_author)
for child2_4 in child2.findall("PubmedData"):
    for child2_5 in child2_4.findall("History
        /PubMedPubDate/[@PubStatus='entrez']"):
        :
        temp_pubdate = child2_5[0].text+"-"+
            child2_5[1].text+"-"+child2_5[2].
            text
    childFinale = child2_4.find("
        ArticleIdList/ArticleId/[@IdType='doi
        ']")
    if (childFinale != None):
        temp_doi = childFinale.text
        print("[ "+str(i)+" ] - ",temp_doi) #
            Debug dell'articolo recuperato

articolo = Articolo (temp_title ,
    autori_vector , temp_pubdate , temp_pubmedid
    , temp_doi ,temp_abstract , temp_jtitle)
articoli.append(articolo)
return articoli

```



## 4.2 Flusso di esecuzione del sistema di apprendimento sugli articoli

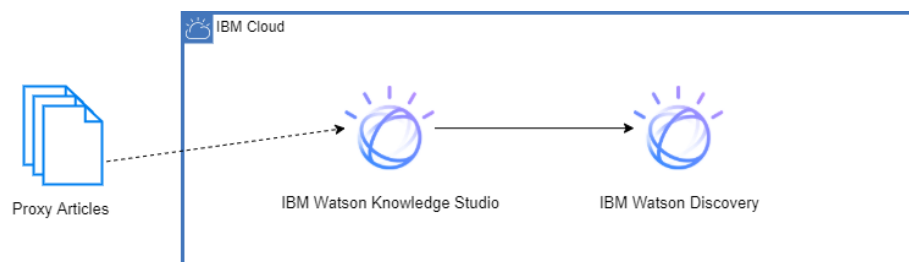


Figura 7: Diagramma di flusso dell'esecuzione

Osserviamo innanzitutto che all'interno di IBM Watson Knowledge Studio vengono memorizzati direttamente gli articoli completi. In particolare, momentaneamente gli articoli sono stati scaricati manualmente sfruttando i Proxy di diverse università campane, quali l'Università degli Studi di Napoli Federico II, l'Università degli Studi della Campania Luigi Vanvitelli e l'Università degli Studi di Salerno.

## 4.3 IBM Watson Knowledge Studio

### 4.3.1 Import degli Articoli

Come anticipato, inizialmente è stato utilizzato un set costituito da soli 8 articoli scaricati manualmente dalle rispettive riviste, attraverso i DOI corrispondenti memorizzati all'interno del database DB2.

In particolare, nella seguente immagine è possibile osservare gli articoli memorizzati all'interno di IBM Watson Knowledge Studio:

Name	Documents
All	7
<a href="#">quinn2009.pdf_set</a>	1
<a href="#">wick2016.pdf_set</a>	1
<a href="#">nabors2011.pdf_set</a>	1
<a href="#">grossmann2012-2.pdf_set</a>	1
<a href="#">beauchesne2010.pdf_set</a>	1
<a href="#">ahluwalia2010.pdf_set</a>	1
<a href="#">wakabayashi2018.pdf_set</a>	1
Items per page: 10 ▾	1-8 of 8 items

Figura 8: Primo set di articoli caricati su IBM Watson Knowledge Studio

### 4.3.2 Creazione del Modello Entità-Relazione

Per realizzare il modello di NLP, innanzitutto, è stato necessario definire le Entità e le Relazioni che dovranno essere annotate all'interno degli articoli.

A seguito di un primo confronto con gli esperti di dominio, essi hanno fornito un primo insieme di entità e relazioni sufficienti per effettuare un'operazione di pre-annotation<sup>3</sup>. In particolare, in seguito a tale primo colloquio, sono state definite le seguenti relazioni e le seguenti entità (evidenziate in grassetto), con rispettivi sinonimi, abbreviazioni e termini appartenenti a quella data entità:

---

<sup>3</sup>Un pre-annotator è un dizionario di Knowledge Studio, un modello Rule-Based oppure un modello di machine learning che può essere eseguito per annotare automaticamente un testo.

# Entità

## TUMOR

- Tumor
- Neoplasm
- Malignancy
- Malignant glioma
- Recurrent malignant glioma
- Progressive malignant glioma
- Oligodendroglioma
- Astrocytoma
- Anaplastic glioma
- Anaplastic astrocytoma
- HGG
- GBM
- Progressive Glioblastoma multiforme
- Recurrent Glioblastoma multiforme
- High grade glioma
- WHO grade IV

## Molecular classification

- 1p/19q codeletion
- IDH1 or IDH2 mutation
- TERT mutation
- EGFR amplicons
- MGMT
- MGMT status
- PTEN mutation
- ATRT
- Peroxisome proliferator-activated receptor gamma

## TRIAL

- Clinical trial registry
- Registry
- Phase I
- Phase II
- Phase III
- Study design
- Open-label
- Exclusion criteria
- Inclusion criteria

## CHEMOTHERAPY

- Chemotherapy
- Nimustine
- Temozolomide
- Carboplatin
- Procarbazine
- Lomustine
- Vincristine
- Carmustine
- Verubulin
- Etoposide
- Cyclophosphamide
- Ifosfamide
- Teniposide

- Interferon
- Convention-enhanced delivery
- CED
- Gliadel wafers
- Carmustine wafers
- Chemotherapy delivery
- Intravenous delivery
- Intra-arterial delivery

#### **TARGET THERAPY**

- Target therapy
- Anti-VEGF
- Bevacizumab
- Anti VEGFR
- Sorafenib
- Anti-EGFR
- erlotinib
- HIV protease inhibitors
- Ritonavir
- Lopinavir

#### **RADIOTHERAPY**

- Radiotherapy
- Radiation therapy
- Irradiation
- Re-irradiation
- Gamma Knife
- Fractionated radiation therapy
- Unfractionated radiation therapy
- Stereotactic radiosurgery

#### **SURGERY**

- Surgery
- Biopsy
- Stereotactic neurosurgery
- Neuronavigation
- Inoperable
- Operable
- Gross total tumor resection
- Complete tumor resection
- Partial resection
- Residual tumor

#### **MAGNETIC RESONANCE IMAGING**

- Magnetic Resonance Imaging
- MRI
- Enhancement
- Contrast-enhancement
- Delayed contrast extravasation MRI
- Treatment Response Assessment Maps
- TRAMs
- Perfusion imaging
- Dynamic Susceptibility Contrast-enhanced

DSC  
Dynamic Contrast-Enhanced  
DCE  
Susceptibility Weighted Imaging

#### **RESPONSE TO THERAPY**

Survival  
Overall survival  
OS  
Progression free survival  
PFS  
Karnofsky status  
Life quality  
Objective response rate  
ORR

#### **ADVERSE EFFECT**

Collateral effect  
Adverse effect  
Adverse reaction  
Secondary reaction  
Dosage  
Toxicity  
Efficacy  
Tolerance  
Safety

#### **RESPONSE ASSESSMENT**

Progression  
True progression  
Regression  
Complete response  
Partial response  
Stable disease  
Pseudoprogression  
Pseudoresponse  
Radionecrosis  
Cerebral radiation necrosis  
Response assessment in neuro-oncology  
RANO criteria  
iRANO criteria

## **Relazioni**

Tumor – molecular classification  
Tumor – chemotherapy  
Tumor - TARGET THERAPY  
Tumor- radiotherapy  
Tumor- surgery  
Tumor- Magnetic Resonance Imaging  
Tumor - trial  
Trial – chemotherapy  
Trial- radiotherapy  
Trial- target therapy  
Trial-surgery

chemotherapy - response to therapy  
chemotherapy – response assessment  
chemotherapy – adverse effect  
target therapy – response to therapy  
target therapy - response assessment  
target therapy - adverse effect  
radiotherapy- response assessment  
radiotherapy – adverse effect  
radiotherapy – response to therapy  
surgery-response assessment  
surgery– adverse effect  
surgery – response to therapy

## **Sinonimi e abbreviazioni**

Neoplasm - malignancy - tumor  
Glioblastoma - GBM - High grade glioma - HGG  
Therapy - treatment  
HIV protease inhibitors - HIVPI  
Radiotherapy - irradiation - radiation therapy - RT  
Gross total tumor resection - Complete tumor resection  
Tolerance - tolerability  
Overall Survival - OS  
Progression free survival - PFS  
Karnofsky status – KS  
Objective response rate - ORR  
Temozolomide - TMZ  
Nimustine - INN  
Interferon – IFN  
Gliadel – carmustine  
Recurrent-progressive  
CED-convection enhanced delivery  
Response assessment in neuro-oncology criteria – RANO  
Magnetic Resonance Imaging - MRI  
Dynamic Susceptibility Contrast-enhanced – DSC  
Dynamic Contrast-Enhanced – DCE  
Treatment Response Assessment Maps - TRAMs

Definito il modello, è stato definito un Dizionario tale che associasse i termini specificati nell'elenco ricevuto dagli esperti di dominio alle rispettive entità.

**Create Dictionary**

**Response assessment**  
Language: English | 13 entries

Entity type: Response\_Assessment Rule class: None

Add Entry Upload Download

Q Enter text to filter

Lemma	Surface Forms	Part of Speech	Action
<input type="checkbox"/> iRANO criteria	iRANO criteria, iRANO Criteria	Unknown	Edit Delete
<input type="checkbox"/> RANO criteria	RANO criteria, RANO Criteria	Unknown	Edit Delete
<input type="checkbox"/> RANO	Response assessment in neuro-oncology, RANO, response assessment in neuro-oncology, Response Assessment in Neuro-oncology, response Assessment in neuro-oncology, response Assessment in Neuro-oncology, Response Assessment in neuro-oncology, Response assessment in Neuro-oncology, response Assessment in Neuro-oncology	Unknown	Edit Delete
<input type="checkbox"/> Cerebral radiation necrosis	Cerebral radiation necrosis, cerebral radiation necrosis, cerebral Radiation necrosis, cerebral radiation Necrosis, Cerebral Radiation necrosis, Cerebral radiation Necrosis, cerebral Radiation Necrosis, Cerebral Radiation Necrosis	Unknown	Edit Delete
<input type="checkbox"/> Radionecrosis	Radionecrosis, radionecrosis	Unknown	Edit Delete
<input type="checkbox"/> Pseudoprogression	Pseudoprogression, pseudoprogression	Unknown	Edit Delete

Figura 9: Dizionario definito

Una volta prodotto il Dizionario, è stata processata la pre-annotazione dei testi precedentemente caricati.

Di seguito è possibile osservare alcuni esempi di risultati ottenuti sui 10 testi:

8 Progressive disease (PD) at 10 weeks was reconfirmed at 18 weeks to distinguish pseudoprogression.

9 **Progression**-free survival (**PFS**), excluding cases of confirmed pseudoprogression, was assessed (post-hoc/exploratory).

10 **Tumor** progression patterns were determined at each disease assessment/PD (prespecified/exploratory).

11 Results.

12 Of patients with PD in the bevacizumab and placebo arms, 143/354 (40.4%) and 155/387 (40.1%), respectively, had PD due to contrast-enhancing lesions, and 51/354 (14.4%) and 53/387 (13.7%) had PD due to nonenhancing lesions.

13 Of all patients in the bevacizumab arm (n=458), 2.2% had confirmed pseudoprogression versus 9.3% in the placebo arm (n=463).

14 Baseline characteristics did not differ between patients with/without pseudoprogression (including for MGMT status).

15 Excluding confirmed pseudoprogression, **PFS** (hazard ratio: 0.65, 95% CI: 0.56-0.75; P = .0001, bevacizumab vs placebo) was comparable to the intent-to-treat population.

16 At PD, most patients had the same tumor focus (local/multifocal, .84%) and infiltrative profile (.88%) as at baseline; no shift to a diffuse or multifocal phenotype was observed.

17 Conclusions.

18 **Pseudoprogression** complicated progression assessment in a small but relevant number of patients but had negligible impact on **PFS**.

19 **Bevacizumab** did not appear to adversely impact tumor progression patterns.

20 Keywords: glioblastoma, pseudoprogression, response, tumor pattern.

21 **Glioblastoma** is an aggressive and invasive brain malignancy with poor prognosis.<sup>1</sup> Standard-of-care therapy for newly diagnosed glioblastoma comprises concurrent radiotherapy and temozolomide (**RT**, **TMZ**), followed by adjuvant **TMZ**.<sup>2</sup> In the randomized

**Type** Subtype Role

**Disease**

**Response**

**Therapy**

**Trial**

Figura 10: Esempio di annotazione

quinn2009.pdf

1 556Neuro-Oncology JULY 2006 Neuro-Oncology This phase 1 clinical trial conducted with patients who had recurrent or progressive malignant glioma (MG) was designed to determine the maximum tolerated dose (MTD) and toxicity of three different 5-day dosing regimens of temozolomide (TMZ) in combination with O6-benzylguanine (O6-BG).

2 Both TMZ and O6-BG were administered on days 1–5 of a 28-day treatment cycle.

3 A bolus infusion of O6-BG was administered at 120 mg/m2 over 1 h on days 1, 3, and 5, along with a continuous infusion of O6-BG at 30 mg/m2/day.

4 TMZ was administered at the end of the first bolus infusion of O6-BG and then every 24 h for 5 days during the continuous infusion of O6-BG.

5 Patients were accrued to one of three 5-day dosing regimens of TMZ.

6 Twenty-nine patients were enrolled into this study.

7 The dose-limiting toxicities (DLTs) were grade 4 neutropenia, leukopenia, and thrombocytopenia.

8 The MTD for TMZ for the three different 5-day dosing schedules was determined as follows: schedule 1, 200 mg/m2 on day 1 and 50 mg/m2/day on days 2–5; schedule 2, 50 mg/m2/day on days 1–5; and schedule 3, 50 mg/m2/day on days 1–5 while receiving pegfilgrastim.

9 Thus, the 5-day TMZ dosing schedule that Phase I trial of temozolomide plus O6-benzylguanine 5-day regimen with recurrent malignant glioma Jennifer A. Quinn, Sara Xiaoyin Jiang, David A. Reardon, Annick Desjardins, James J. Vredenburg, Jeremy N. Rich, Sridharan Gururangan, Allan H. Friedman, Darell D. Bigner, John H. Sampson, Roger E. McLendon, James E. Herndon, Jr., Amy Walker, and Henry S. Friedman Departments of Surgery (J.A.Q., S.X.J., D.A.R., A.D., J.J.V., J.N.R., S.G., A.H.F., D.D.B., J.H.S.,

Type	Subtype	Role
Disease		
Response		
Therapy		
Trial		

Figura 11: Esempio di annotazione

Back to Annotations

Open document list

View Details

Replace

Concordance

Attribute View

Save

Annotation

Relation

Conference

20 3 Neurooncol 123 combination of verubulin at the previously determined single-agent maximum tolerated dose of 3.3 mg/m2 with carboplatin in patients with recurrent/refractory GBM is safe and well tolerated.

21 In this patient population with a highly vascularized tumor, no cerebral hemorrhage was observed.

22 Keywords Glioblastoma multiforme Verubulin Carboplatin Vascular disrupting agent Introduction Glioblastoma multiforme (GBM) is one of the most aggressive central nervous system (CNS) malignancies, affecting more than 20,000 people in the United States annually [1].

23 Surgery followed by concurrent daily temozolomide and cranial irradiation followed by at least 6 cycles of post-radiation temozolomide has become the standard of care for these patients.

24 Despite this aggressive therapy, median survival remains approximately 15–18 months for newly diagnosed GBM and 2-year overall survival rates remain only approximately 25 % [2, 3].

25 While there have also been advances in treatment options for recurrent disease, the outcomes in this patient population remain poor.

26 Prior to the availability of bevacizumab therapy, 6-month progression-free survival (PFS) and overall survival (OS) rates for recurrent GBM were approximately 16 and 55 %, respectively, and median PFS and OS were 8 and 30 weeks, respectively [4].

27 With the advent of bevacizumab treatment, improvements have been seen in 6-month and median PFS rates, although OS rates remain disappointing [5].

28 A number of factors likely contribute to the poor efficacy of chemotherapy for recurrent GBM, including poor penetration of many chemotherapeutic agents into the CNS because of the blood–brain barrier, as well as abnormal vasculature and high interstitial pressures.

Type	Subtype	Role
Disease		
Response		
Therapy		
Trial		

Figura 12: Esempio di annotazione

#### 4.4 Valutazione del modello

Per valutare il modello prodotto, è stato inserito un ulteriore pull di 40 articoli presenti nel database ed inseriti all'interno di Knowledge Studio. Una volta caricati, è stata effettuata la pre-annotazione su di essi e, una volta terminata, si è proceduto con l'addestramento del modello e la sua valutazione; per effettuare tale procedura, si è proceduto con uno split dell'intero set di documenti nel seguente modo:

- Training Set: 70% dell'intero set di documenti
- Test Set: 23% dell'intero set di documenti
- Blind Set: 7% dell'intero set di documenti

23



☒ Create new sets by splitting the selected document sets

---

**Ratio**  
*Enter the percentage of documents to include in each set.*

70	Training Set (70% Recommended)
23	Test Set (23% Recommended)
7	Blind Set (7% Recommended)

---

Figura 13: Suddivisione del set di documenti

Si osservi che il numero di document sets differisce dal numero totale di documenti. Ciò è dovuto alle limitazioni imposte da Knowledge Studio sui documenti caricati: infatti, il tool non effettua il caricamento di documenti di dimensione maggiore di 5MB oppure eccedenti i 40'000 caratteri. Per rientrare nei limiti imposti, abbiamo operato la suddivisione di un articolo in più parti, e abbiamo successivamente caricato le parti come un unico document set.

Nello specifico, il Training Set viene utilizzato per addestrare il modello sulle annotazioni effettuate, il Test Set viene utilizzato dal modello (dopo essersi addestrato tramite il Training) per effettuare una nuova annotazione sui documenti che contiene per poi valutare gli errori effettuati essendo questi ultimi già stati pre-annotati in precedenza; il Blind Set, infine, viene utilizzato per testare periodicamente il sistema a valle di diverse iterazioni di testing e miglioramento. Rispetto al test set, il blind set deve contenere elementi che non sono stati visti dagli utenti coinvolti nell'annotazione. Quest'aspetto permette di distinguere il blind set dal test set: quest'ultimo, infatti, contiene documenti che vengono periodicamente esaminati dagli utenti coinvolti nel processo di annotation, e questo tipo di conoscenza può contribuire ad alterare impropriamente il valore di precision. Avere un insieme di documenti non visualizzato da tali utenti permette di soffermarsi solo sugli score di alto livello, senza pilotare il miglioramento del modello sulla base della conoscenza che gli utenti hanno del test set.

Prodotta la suddivisione, si è proceduto con l'addestramento del sistema e la sua valutazione tramite le metriche di accuracy, precision, recall e F1-score: l'intera operazione ha richiesto un tempo di 18'14".

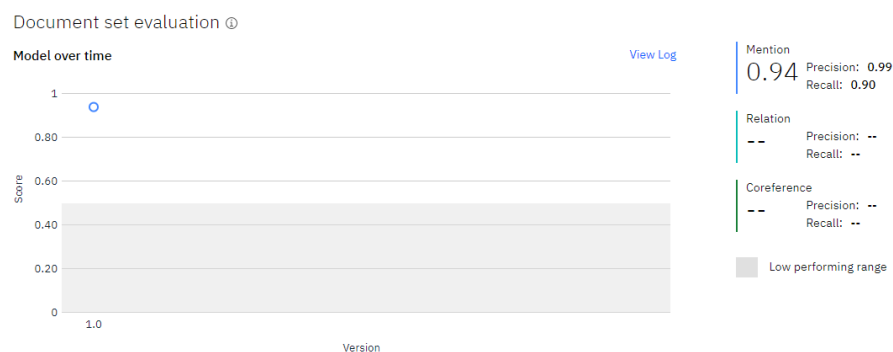


Figura 14: Accuracy, Precision e Recall del modello

Osserviamo che il dizionario è stato popolato solo da termini appartenenti a determinate entità, la valutazione del modello e delle relative metriche è stata effettuata solo in base all'annotazione delle entità, in quanto la pre-annotazione tramite dizionario non consente di individuare automaticamente all'interno dei documenti le relazioni da noi specificate. Il sistema restituisce nel dettaglio i valori delle metriche per ogni entità definita nel modello.

Test Set Summary Entity Types Relation Types Coreferenced Mentions

Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
Adverse_Effect	0.86	1	0.76	7% (165/2349)	0% (165/76780)	77% (17/22)
Chemotherapy	0.59	0.97	0.42	18% (412/2349)	1% (412/76780)	86% (19/22)
Magnetic_Resonance_Imagi...	0.91	0.98	0.85	4% (86/2349)	0% (86/76780)	55% (12/22)
Molecular_classification	0.9	0.96	0.85	4% (87/2349)	0% (87/76780)	41% (9/22)
Radiotherapy	0.93	0.98	0.88	13% (306/2349)	0% (306/76780)	73% (16/22)
Response	0.98	0.98	0.98	16% (370/2349)	0% (370/76780)	95% (21/22)
Response_Assessment	0.83	1	0.71	4% (93/2349)	0% (93/76780)	91% (20/22)
Surgery	0.91	1	0.83	3% (80/2349)	0% (80/76780)	73% (16/22)
Target_Therapy	0.71	1	0.56	1% (18/2349)	0% (18/76780)	36% (8/22)
Trial	0.76	0.87	0.68	6% (149/2349)	0% (149/76780)	73% (16/22)
Tumor	0.91	0.98	0.85	25% (583/2349)	1% (583/76780)	95% (21/22)
Overall Statistics	0.86	0.98	0.77	100% (2349/2349)	3% (2349/76780)	100% (22/22)

Figura 15: Valori di Accuracy, Precision, Recall e F1-Score delle entità

Infine, è possibile valutare l'accuracy di ogni singola entità tramite un istogramma che riassume i singoli punteggi ottenuti.

Current version insights

Mention breakdown ⓘ Detailed Statistics

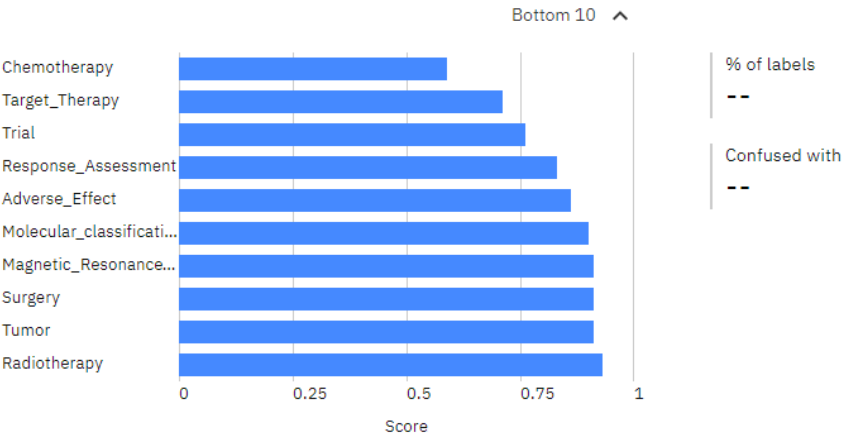


Figura 16: Score di accuracy ottenuto dalle entità delle entità

## 5 Apprendimento sulle Risonanze Magnetiche tramite Visual Recognition

### 5.1 Flusso di esecuzione del sistema di apprendimento sulle MRI

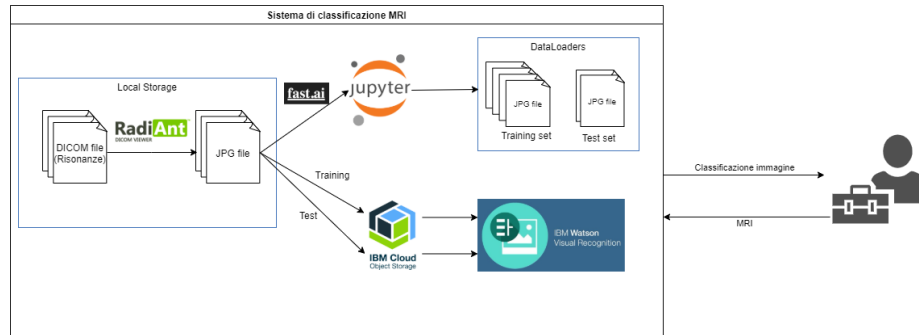


Figura 17: Diagramma di flusso dell'esecuzione

In un primo momento, i file DICOM sono stati sottoposti ad una fase di pre-processing: essi sono stati aperti tramite il tool *RadiAnt DICOM Viewer* e convertiti in formato JPG.

Dopo la fase di pre-processing, le risonanze sono state utilizzate per la realizzazione di due reti differenti:

1. il tool di Visual Recognition di Watson Studio.
2. una CNN implementata nella libreria *fastai*;

Quest'ultima è stata usata in fase preliminare e come metro di paragone con Watson Visual Recognition. La sua implementazione e utilizzo sono specificati nella Sezione C dell'Appendice del presente documento.

### 5.2 Visual Recognition di Watson Studio

L'addestramento del classificatore attraverso Watson Studio è stato addestrato attraverso diversi cicli di raffinamento che analizzeremo nel dettaglio di seguito:

1. Primo Ciclo: il classificatore è stato addestrato fornendogli in ingresso un Training set costituito da un totale di 691 immagini, divise in due diverse classi. La prima è la classe GBM che contiene 489 immagini di glioblastomi multiformi, mentre la seconda è la classe MET che contiene 202 immagini di metastasi. Osserviamo che, in pratica il Training Set adoperato per realizzare il primo ciclo di addestramento è, in pratica, lo stesso Training Set utilizzato per addestrare il classificatore con *fastai*;
2. Secondo Ciclo: il classificatore è stato addestrato fornendogli in ingresso un Training set costituito da un totale di 1'838 immagini, divise in due

diverse classi ed un set di immagini definite come Negative. In particolare, le classi GBM e MET sono rimaste invariate rispetto al primo ciclo di addestramento, mentre le 1'147 immagini Negative sono immagini che non appartengono a nessuna di esse;

3. Terzo Ciclo: il classificatore è stato addestrato fornendogli in ingresso un Training set costituito da un totale di 2'103 immagini, divise in due diverse classi ed un set di immagini definite come Negative. In particolare, le immagini definite come Negative sono rimaste invariate, mentre sono state aggiunte 221 immagini classificate come GBM e 44 immagini classificate, invece, come MET;
4. Quarto Ciclo: il classificatore è stato addestrato fornendogli in ingresso un Training set costituito da un totale di 956 immagini, divise in due diverse classi. In particolare, le immagini appartenenti alle classi GBM e MET sono le stesse utilizzate nel terzo ciclo di addestramento. Sono state, invece, rimosse le immagini definite come Negative.

Al termine di ogni ciclo di addestramento sono state valutate le prestazioni del classificatore in termini di metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score**, attraverso l'utilizzo di un Test Set costituito da un totale di 327 elementi suddivisi in due classi, quali GBM e MET. In particolare, 223 elementi sono classificati come GBM ed i restanti 104 elementi sono, invece, classificati come MET.

### 5.2.1 Implementazione

Innanzitutto, una prima parte di dati etichettati forniti dagli esperti di dominio è stata utilizzata come Training Set con lo scopo di far apprendere il modello.

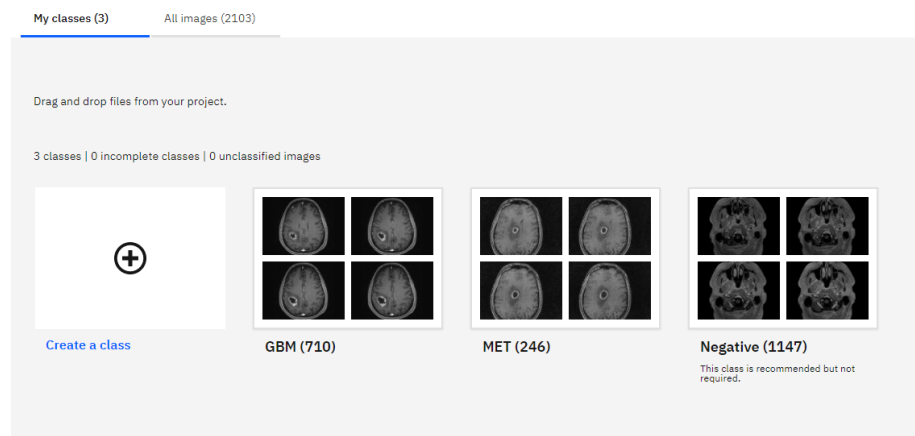


Figura 18: Definizione delle classi su Visual Recognition

Dopodiché, un secondo insieme di dati etichettati dagli esperti di dominio è stato utilizzato come Test Set al fine di poter valutare la metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score**.

In particolare, è stato realizzato un notebook in linguaggio Python che, sfruttando le IBM Cloud API, ci ha consentito di utilizzare il classificatore precedentemente realizzato per effettuare una classificazione del Test Set.

In primo luogo è stata generata un'istanza di **VisualRecognitionV3**, che consente di realizzare il collegamento con l'istanza di Visual Recognition addestrata all'interno di Watson Studio. Ciò è stato realizzato attraverso la funzione *createVRinstance(parameters)*, che riceve in ingresso un insieme di parametri specificati all'interno di un file JSON. In particolare, i campi specificati all'interno del file sono i seguenti:

- url;
- apikey;
- vr\_id: id del classificatore da noi istruito;
- test\_csv\_file: file csv contenente i nomi delle immagini appartenenti al Test Set e le rispettive classificazioni;
- results\_csv\_file: file csv in cui verranno memorizzati i risultati;
- confmatrix\_csv\_file: file csv in cui verrà memorizzata la Confusion Matrix (Matrice di Confusione<sup>4</sup>).

Le informazioni contenute all'interno del file JSON vengono memorizzate all'interno di un oggetto di tipo Parameters:

```
def readParameters(parmsPath):
    #Lettura dei parametri utili da un file json
    vrParmsFile = parmsPath
    parms = ''
    with open(vrParmsFile) as parmFile:
        parms = json.load(parmFile)

    parameters = Parameters(parms['url'], parms['apikey'],
                            parms['vr_id'], parms['test_csv_file'], parms['results_csv_file'],
                            parms['confmatrix_csv_file'],
                            parms['version'])
    json.dumps(parms)

    return parameters
```

---

<sup>4</sup>Nell'ambito dell'Intelligenza artificiale, la matrice di confusione, detta anche tabella di errata classificazione, restituisce una rappresentazione dell'accuratezza di classificazione statistica. Ogni colonna della matrice rappresenta i valori predetti, mentre ogni riga rappresenta i valori reali. L'elemento sulla riga i e sulla colonna j è il numero di casi in cui il classificatore ha classificato la classe "vera" i come classe j. Attraverso questa matrice è osservabile se vi è "confusione" nella classificazione di diverse classi.

I parametri appena acquisiti, come anticipato, vengono utilizzati per la creazione di un'istanza di Visual Recognition, ovvero un'istanza della classe **VisualRecognitionV3**, definita all'interno della libreria **ibm\_watson**. L'implementazione della funzione *createVRinstance(parameters)* è riportata di seguito:

```
def createVRinstance(parameters):
    #Creazione di una istanza di Visual Recognition
    authenticator = IAMAuthenticator(parameters.apiKey)
    visualRecognition = VisualRecognitionV3(
        version = parameters.version,
        authenticator = authenticator
    )
    visualRecognition.set_service_url(parameters.url)

    return visualRecognition
```

Dopo la generazione dell'istanza della classe **VisualRecognitionV3**, è possibile effettuare la classificazione delle immagini contenute all'interno del Test Set. A tal fine, innanzitutto è stata definita la funzione *getVRresponse* che va a classificare una singola immagine restituendo una lista delle classi associate all'immagine stessa, ordinata in base al valore di Confidence (score). La funzione riceve in ingresso 3 parametri, quali l'istanza di visual recognition, l'id del classificatore ed il path dell'immagine da classificare. L'implementazione della funzione è riportata di seguito:

```
def getVRresponse(vr_instance, classifierID, image_path):
    with open(image_path, 'rb') as images_file:
        image_results = vr_instance.classify(
            images_file = images_file,
            threshold='0.6',
            classifier_ids=classifierID).get_result()

    #Ogni singola immagine viene classificata attraverso
    il classificatore da noi istruito
    classList = []
    for classifier in image_results['images'][0]['classifiers']:
        if classifier['classifier_id'] == parameters.vrId:
            :
            classList = classifier['classes']
        break
    #Le classi vengono ordinate in base allo score
    sorted_classList = sorted(classList, key=lambda k: k.get('score', 0), reverse=True)
    return sorted_classList
```

Sfruttando la funzione appena definita, è stato possibile definire un'ulteriore funzione *batchVR*, che consente di classificare un batch di immagini restituendo,

per ciascuna di esse, i risultati ottenuti. In particolare, i path delle immagini appartenenti al batch da classificare vengono specificati all'interno di un file **csv**, insieme alla classe di appartenenza dell'immagine stessa. La funzione riceve in ingresso 3 parametri, quali l'istanza di visual recognition, l'id del classificatore ed il path del file csv. L'implementazione della funzione è riportata di seguito:

```
def batchVR(vr_instance , classifierID , csvfile):
    testClasses=[]
    vrPredictClasses=[]
    vr_predict_confidence=[]
    images=[]
    i=0
    with open(csvfile , 'r') as csvfile:
        csvReader=csv.DictReader(csvfile)
        for row in csvReader:
            testClasses.append(row[ 'class '])
            vr_response = getVRresponse(vr_instance ,
                classifierID ,row[ 'image '])
            vrPredictClasses.append(vr_response[0][ 'class
                '])
            vr_predict_confidence.append(vr_response[0][ '
                score '])
            images.append(row[ 'image '])
            i = i+1
            if(i%250 == 0):
                print("")
                print("Processed " , i , " records")
            if(i%10 == 0):
                sys.stdout.write('.')
        print("")
        print("Finished processing " , i , " records")
    return testClasses , vrPredictClasses ,
        vrPredictConfidence , images
```

### 5.2.2 Risultati ottenuti - Primo Ciclo

Il tempo impiegato per effettuare il Training del classificatore dato il Training Set descritto precedentemente (che consta di 691 immagini) è pari a 36'17".

Le metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score** sono state stimate attraverso l'utilizzo delle funzioni *accuracy\_score* e *classification\_report*, entrambe definite all'interno della libreria **sklearn**.

Di seguito sono riportati i risultati ottenuti:

```
accuracy = accuracy_score(testClasses , vrPredictClasses)
print('Classification Accuracy: ' , accuracy)
```



**Classification Accuracy: 0.6452599388379205**

Figura 19: Accuracy della classificazione ottenuta

```
classification_report(testClasses, vrPredictClasses,
                      labels=labels)
```

	precision	recall	f1-score	support
GBM	0.71	0.80	0.76	223
MET	0.42	0.31	0.36	104
accuracy			0.65	327
macro avg	0.57	0.56	0.56	327
weighted avg	0.62	0.65	0.63	327

Figura 20: Valori di Precision, Recall e F1-Score

Inoltre, è stata generata e memorizzata all'interno di un file **csv** la matrice di confusione attraverso la funzione *confusion\_matrix*, anch'essa definita all'interno della libreria **sklearn**:

```
labels=list(set(testClasses))
vrConfusionMatrix = confusion_matrix(testClasses,
                                     vrPredictClasses, labels)
confmatrix2csv(vrConfusionMatrix, labels, parameters.
               confMatrixCsvFile)
```

La matrice di confusione ottenuta è, quindi, la seguente:

```
df_cm = pd.DataFrame(vrConfMatrix, range(2), range(2))
sn.set(font_scale=1.4). # Label size
sn.heatmap(df_cm, annot=True, annot_kws={"size": 16})
# Font size
plt.show()
```

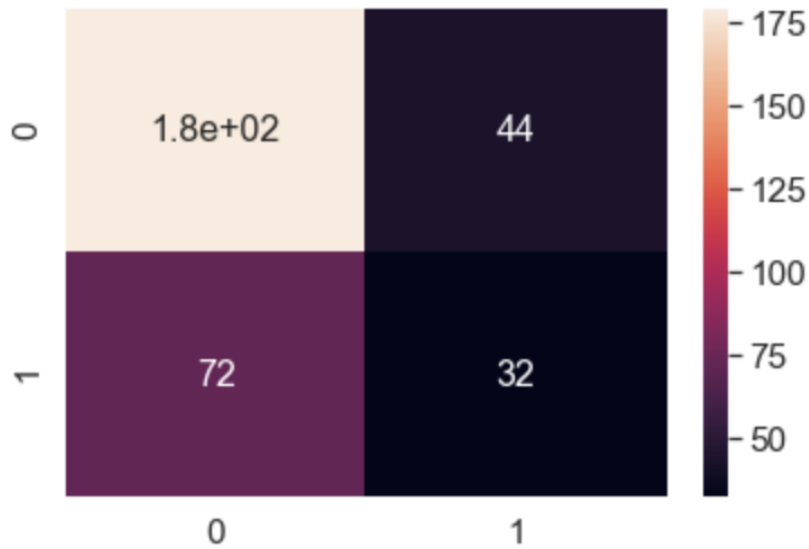


Figura 21: Matrice di Confusione

Infine, tale matrice è stata memorizzata all'interno di un file **csv** attraverso la funzione *confmatrix2csv*, la cui implementazione è riportata di seguito:

```
def confmatrix2csv(conf_matrix, labels, csvfile):
    with open(csvfile, 'w') as csvfile:
        csvWriter = csv.writer(csvfile)
        row=list(labels)
        row.insert(0,"")
        csvWriter.writerow(row)
        for i in range(conf_matrix.shape[0]):
            row=list(conf_matrix[i])
            row.insert(0,labels[i])
            csvWriter.writerow(row)
```

### 5.2.3 Risultati ottenuti - Secondo Ciclo

Il tempo impiegato per effettuare il Training del classificatore dato il Training Set descritto precedentemente (che consta di 1'838 immagini) è pari a 1h56'59". Le metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score** sono state stimate attraverso l'utilizzo delle funzioni *accuracy\_score* e *classification\_report*, entrambe definite all'interno della libreria **sklearn**.

Di seguito sono riportati i risultati ottenuti:

```
accuracy = accuracy_score(testClasses, vrPredictClasses)
print('Classification Accuracy: ', accuracy)
```

**Classification Accuracy: 0.5192307692307693**

Figura 22: Accuracy della classificazione ottenuta

```
classification_report(testClasses, vrPredictClasses,
                      labels=labels)
```

	precision	recall	f1-score	support
GBM	1.00	0.52	0.68	52
micro avg	1.00	0.52	0.68	52
macro avg	1.00	0.52	0.68	52
weighted avg	1.00	0.52	0.68	52

Figura 23: Valori di Precision, Recall e F1-Score

Osserviamo, quindi, che c'è stato un calo dell'accuracy del classificatore di circa il 20% rispetto al primo ciclo di addestramento. Tale diminuzione è una conseguenza del fatto che il classificatore non sia più stato in grado di riconoscere le metastasi (immagini appartenenti alla classe MET). Infatti, tutte le immagini classificate come MET appartenenti al Test Set sono state erroneamente riconosciute come Negative.

Inoltre, è stata generata e memorizzata all'interno di un file **csv** la matrice di confusione attraverso la funzione *confusion\_matrix*, anch'essa definita all'interno della libreria **sklearn**:

```
labels=list(set(testClasses))
vrConfusionMatrix = confusion_matrix(testClasses,
                                     vrPredictClasses, labels)
confmatrix2csv(vrConfusionMatrix, labels, parameters.
               confMatrixCsvFile)
```

La matrice di confusione ottenuta è, quindi, la seguente:

```
df_cm = pd.DataFrame(vrConfMatrix, range(2), range(2))
sn.set(font_scale=1.4). # Label size
sn.heatmap(df_cm, annot=True, annot_kws={"size": 16})
# Font size
plt.show()
```

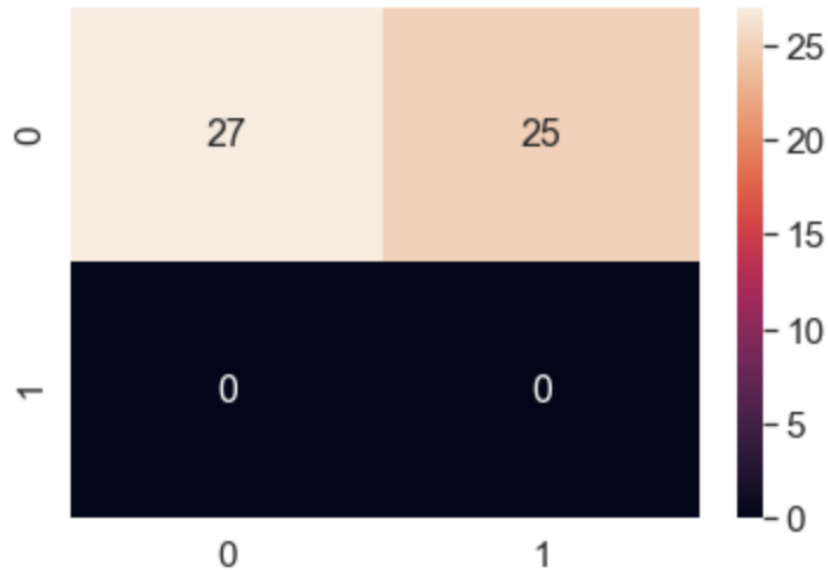


Figura 24: Matrice di Confusione

Dalla matrice è possibile osservare come, appunto, il classificatore non abbia classificato affatto le immagini appartenenti alla classe MET che, invece, sono state riconosciute tutte come Negative.

Infine, tale matrice è stata memorizzata all'interno di un file **csv** attraverso la funzione *confmatrix2csv*.

#### 5.2.4 Risultati ottenuti - Terzo Ciclo

Il tempo impiegato per effettuare il Training del classificatore dato il Training Set descritto precedentemente (che consta di 2'103 immagini) è pari a 2h0'31". Le metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score** sono state stimate attraverso l'utilizzo delle funzioni *accuracy\_score* e *classification\_report*, entrambe definite all'interno della libreria **sklearn**.

Di seguito sono riportati i risultati ottenuti:

```
accuracy = accuracy_score(testClasses , vrPredictClasses)
print('Classification Accuracy: ', accuracy)
```

**Classification Accuracy: 0.826530612244898**

Figura 25: Accuracy della classificazione ottenuta

```
classification_report(testClasses , vrPredictClasses ,
                      labels=labels)
```

	precision	recall	f1-score	support
MET	0.03	0.20	0.06	5
GBM	0.98	0.84	0.90	191
accuracy			0.83	196
macro avg	0.50	0.52	0.48	196
weighted avg	0.95	0.83	0.88	196

Figura 26: Valori di Precision, Recall e F1-Score

Inoltre, è stata generata e memorizzata all'interno di un file **csv** la matrice di confusione attraverso la funzione *confusion\_matrix*, anch'essa definita all'interno della libreria **sklearn**:

```
labels=list(set(testClasses))
vrConfusionMatrix = confusion_matrix(testClasses ,
    vrPredictClasses , labels)
confmatrix2csv(vrConfusionMatrix , labels , parameters .
    confMatrixCsvFile)
```

La matrice di confusione ottenuta è, quindi, la seguente:

```
df_cm = pd.DataFrame(vrConfMatrix , range(2) , range(2))
sn.set(font_scale=1.4). # Label size
sn.heatmap(df_cm , annot=True , annot_kws={"size": 16})
# Font size
plt.show()
```

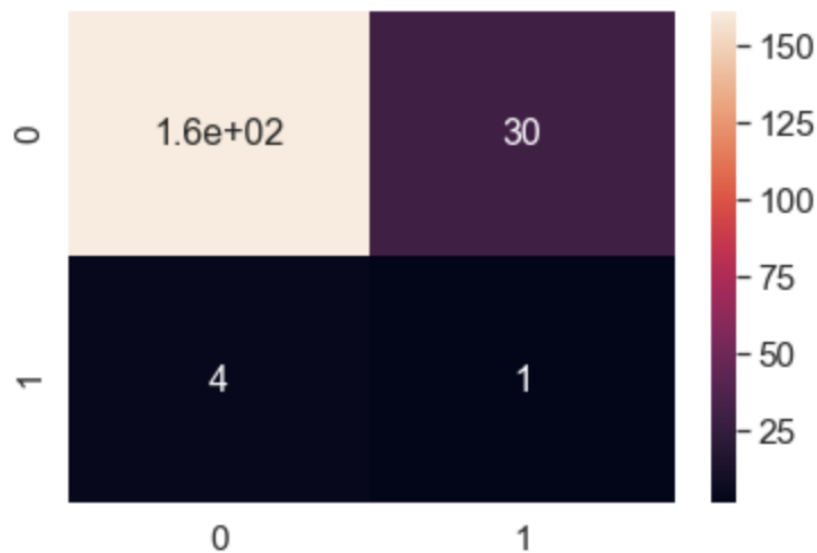


Figura 27: Matrice di Confusione

Infine, tale matrice è stata memorizzata all'interno di un file **csv** attraverso la funzione *confmatrix2csv*.

### 5.2.5 Risultati ottenuti - Quarto Ciclo

Il tempo impiegato per effettuare il Training del classificatore dato il Training Set descritto precedentemente (che consta di 956 immagini) è pari a 1h04'28". Le metriche di **Accuracy**, **Precision**, **Recall** ed **F1-score** sono state stimate attraverso l'utilizzo delle funzioni *accuracy\_score* e *classification\_report*, entrambe definite all'interno della libreria **sklearn**.

Di seguito sono riportati i risultati ottenuti:

```
accuracy = accuracy_score(testClasses , vrPredictClasses)
print('Classification Accuracy: ', accuracy)
```

**Classification Accuracy: 0.7706422018348624**

Figura 28: Accuracy della classificazione ottenuta

```
classification_report(testClasses , vrPredictClasses ,
                      labels=labels)
```

	precision	recall	f1-score	support
GBM	0.78	0.92	0.85	223
MET	0.72	0.45	0.56	104
accuracy			0.77	327
macro avg	0.75	0.69	0.70	327
weighted avg	0.76	0.77	0.75	327

Figura 29: Valori di Precision, Recall e F1-Score

Inoltre, è stata generata e memorizzata all'interno di un file **csv** la matrice di confusione attraverso la funzione *confusion\_matrix*, anch'essa definita all'interno della libreria **sklearn**:

```
labels=list(set(testClasses))
vrConfusionMatrix = confusion_matrix(testClasses ,
                                     vrPredictClasses , labels)
confmatrix2csv(vrConfusionMatrix , labels , parameters.
               confMatrixCsvFile)
```

La matrice di confusione ottenuta è, quindi, la seguente:

```
df_cm = pd.DataFrame(vrConfMatrix, range(2), range(2))
sn.set(font_scale=1.4). # Label size
sn.heatmap(df_cm, annot=True, annot_kws={"size": 16})
# Font size
plt.show()
```

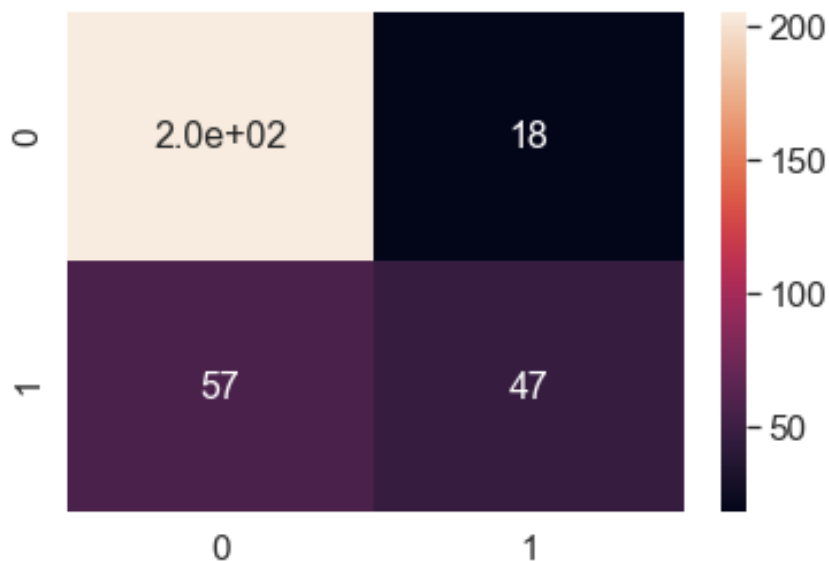


Figura 30: Matrice di Confusione

Infine, tale matrice è stata memorizzata all'interno di un file **csv** attraverso la funzione *confmatrix2csv*:

## confmatrix\_csv

	GBM	MET
GBM	205	18
MET	57	47

Figura 31: File CSV contenente la Matrice di Confusione

## 6 Considerazioni e Sviluppi Futuri

Il modello di Visual Recognition fornito da IBM risulta essere molto efficiente e rapido nell'addestramento su grosse moli di dati di training; dopo un confronto con gli esperti di dominio, è emerso che il lasso di tempo per addestrare il classificatore sul set di 2'103 (ovvero la cardinalità del dataset maggiore su cui abbiamo lavorato) pari a circa 2h risulta essere inferiore al tempo che impiegerebbe un essere umano ad analizzare lo stesso campione di dati.

Infatti, come è possibile osservare dal seguente grafico, i tempi impiegati per effettuare l'addestramento del classificatore sono pressoché lineari con l'aumentare dei dati presenti all'interno del Training Set:

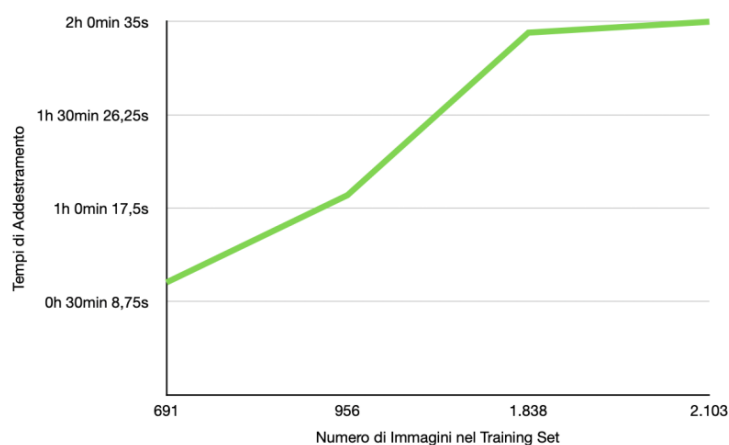


Figura 32: Grafico sull'andamento dei tempi di addestramento del classificatore

Le criticità emerse durante l'addestramento e la valutazione del modello riguardano principalmente l'utilizzo di immagini di risonanze "neutre" (inserite nel classificatore come *Negative*) che inducevano il modello a classificare erroneamente diverse immagini inerenti alle Metastasi: queste venivano identificate come Negativi, quindi non appartenenti alla classe Glioblastoma (*GBM*) né alla classe Metastasi (*MET*), e dunque non venivano utilizzate per il calcolo delle metriche valutative del modello. Inoltre, la poca trasparenza nella documentazione fornita da IBM sugli algoritmi di Deep Learning usati, non ci ha potuto permettere di usare una strategia precisa su come procedere per il superamento della criticità sopracitata se non tramite un procedimento di "trial and error", ovvero generando un modello che presentasse solamente immagini appartenenti alle classi GBM e MET rimuovendo i Negative. Seppur inizialmente questo tipo di modello non ci aveva fornito risultati soddisfacenti, soprattutto per la classificazione delle Metastasi, aumentando la cardinalità del training set al numero finale di 956 immagini complessive (710 GBM, 246 MET) il modello ha mostrato risultati soddisfacenti sulla classificazione di entrambe le classi.

Per poter migliorare ancora il modello, bisognerebbe reperire ulteriori risonanze



di Metastasi con cui indurre nuovi addestramenti incrementali del sistema, oltre che valutarne l'efficienza fornendogli set di risonanze prese da più piani (tutte le immagini fornite erano risonanze T1 con mezzo di contrasto sullo stesso piano) e con spessori diversi.

Il modello di annotazione dei documenti creato su Watson Knowledge Studio ha permesso un'individuazione facile dei termini necessari per l'analisi e lo studio dei testi inerenti al Glioblastoma; tuttavia, la difficoltà nel reperire tutti gli articoli individuati dalla query elaborata insieme agli esperti di dominio e la limitazione di Knowledge Studio nel poter operare su file contenenti un massimo di 40'000 caratteri ha richiesto un lungo e dispendioso lavoro di pre-processing che ha inficiato la possibilità di poter operare insieme agli esperti di dominio per un tagging manuale delle entità e delle relazioni presenti all'interno dei testi utilizzati al fine di migliorare ulteriormente le performance del sistema (nonostante risultino essere elevate anche con la sola pre-annotazione) e individuare termini non presenti nel dizionario fornito. Una più profonda fase di tagging permetterebbe di poter elaborare una futura strategia di analisi dei testi in modo da poter superare i metodi precedentemente utilizzati in altri studi, e forniti a noi come punto di partenza per la creazione del prototipo esposto in questa documentazione.

# Appendice

## A RadiAnt

RadiAnt DICOM Viewer è un visualizzatore di file DICOM sviluppato dalla compagnia Medixant. Il suo primo rilascio, in versione Beta, risale al 2010 con la versione 0.1, ad oggi è arrivato alla versione 2020.1 rilasciato il 6 Aprile 2020. Attraverso un'interfaccia grafica user-friendly, RadiAnt fornisce strumenti per lo studio delle immagini DICOM quali:

- Zoom e spostamento;
- Regolazione della luminosità e del contrasto;
- Rotazione e capovolgimento delle immagini;
- Conversione in formati BMP e JPG.

Il software supporta l'apertura di file ottenuti da differenti modalità di immagini quali:

- Radiografia Digitale (CR, DX)
- Mammografia (MG)
- Tomografia Computerizzata (CT)
- Risonanza Magnetica (MR)
- Tomografia ad Emissione di Positroni PET-CT (PT)
- Ecografia (US)
- Angiografia Digitale (XA)
- Gamma Camera, Medicina Nucleare (NM)
- Immagini secondarie e immagini digitalizzate (SC)
- Structured Reports (SR)

RadiAnt fornisce numerose funzionalità, tra cui alcune da noi sfruttate, quali:

- **Confronto di serie o studi differenti**  
È possibile aprire contemporaneamente più serie di un determinato studio oppure di studi differenti nelle stesse o in diverse finestre al fine di effettuare una comparazione fra essi.
- **Ricostruzioni Multiplanari**  
Strumento in gradi di ricostruire immagini su piani ortogonali differenti (coronali, sagittali, obliqui, assiali).

- **Esportazione di files DICOM come immagini e filmati**  
È possibile esportare i file DICOM aperti sia come immagini nei formati JPG e BMP sia come video in formato WMV (Windows Media Video) e MP4.
- **Rendering del Volume 3D**  
Strumento che consente di visualizzare grandi volumi di dati generati dagli scanner CT/MR in uno spazio tridimensionale.

### A.1 Pre-processing effettuato tramite RadiAnt

L'utilizzo di RadiAnt è risultato fondamentale nella fase di pre-processing dei file DICOM prima che venissero dati in pasto al classificatore dell'istanza di IBM Visual Recognition; il software permette la lettura di intere cartelle di file DICOM, unendole in singole sequenze.

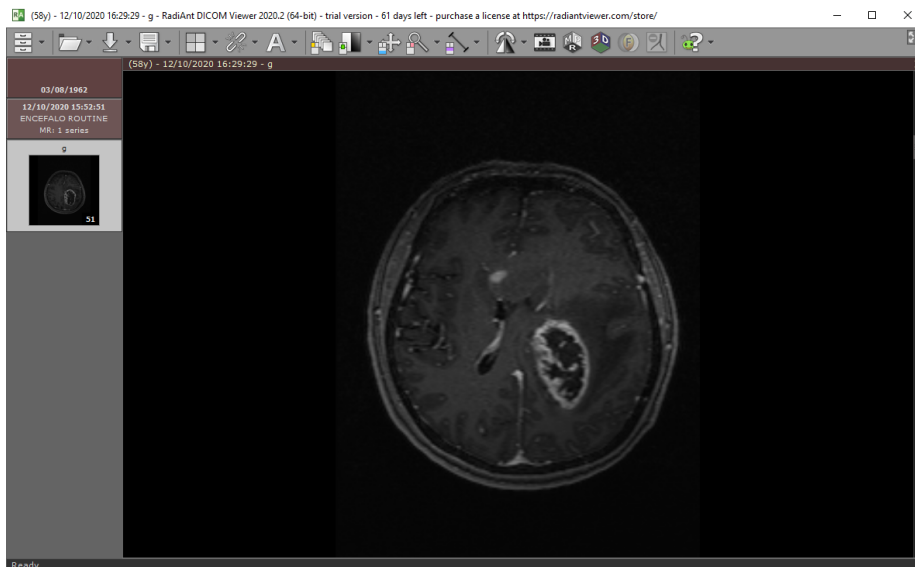


Figura 33: Apertura di una singola cartella contenente una sequenza composta da 52 file DICOM

È anche possibile caricare nel software più cartelle contemporaneamente in modo da poter visualizzare contemporaneamente più sequenze.

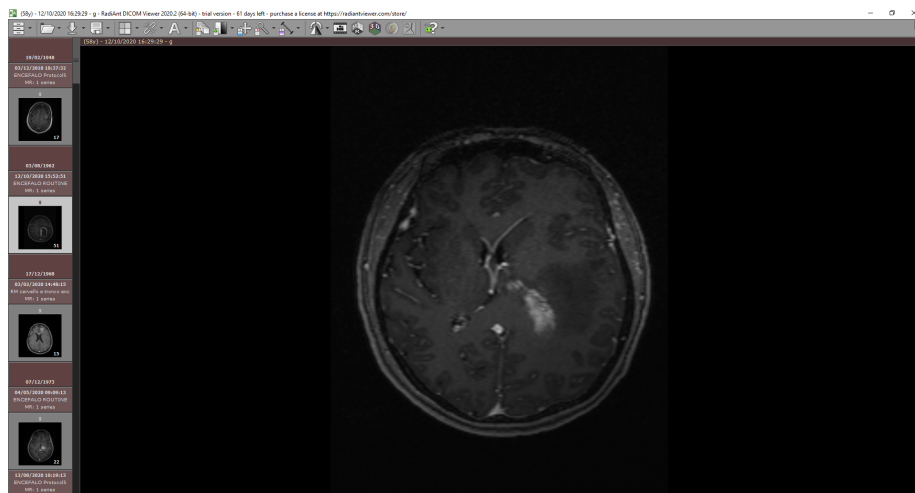


Figura 34: Apertura di più sequenze

Una volta aperte le MRI di nostro interesse, è stata utilizzata la funzionalità di conversione in JPG.

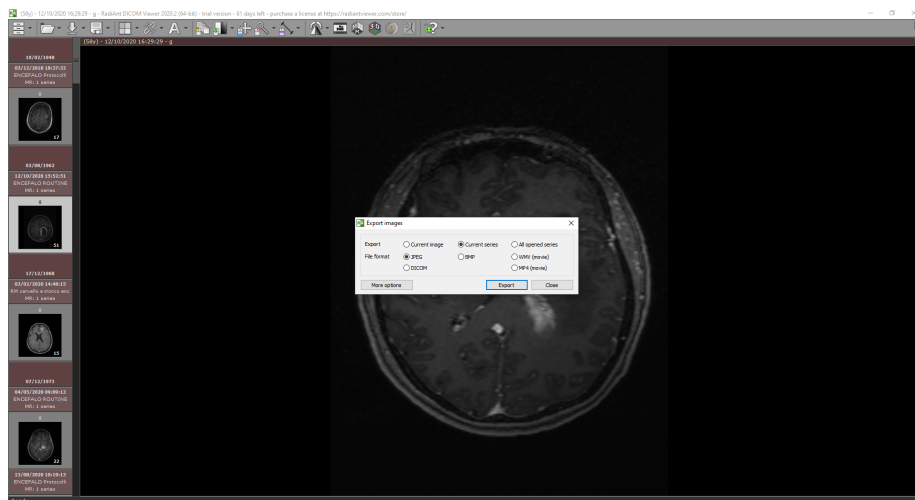


Figura 35: Conversione ed Export in JPG

Le immagini delle risonanze in JPG sono state compresse in un file ZIP in modo da poter caricare l'intero set di sequenze nel Visual Recognition.

## B Evaluation Metrics

### B.1 Accuracy

L'Accuracy, nell'ambito del Machine Learning, può essere definita come il rapporto tra il numero di elementi predetti correttamente ed il supporto (ovvero il numero totale di elementi).

### B.2 Precision e Recall

In un processo di classificazione, Precision e Recall vengono calcolate in termini di Veri Positivi, Falsi Positivi, Veri Negativi e Falsi Negativi, effettuando un confronto tra la classe assegnata all'oggetto dal classificatore stesso e quella a cui realmente esso appartiene.

Formalmente, Precision e Recall sono definite come:

$$Precision = \frac{Veri\ Positivi}{Veri\ Positivi + Falsi\ Positivi}$$

$$Recall = \frac{Veri\ Positivi}{Veri\ Positivi + Falsi\ Negativi}$$

Ad esempio, si può considerare la seguente figura:

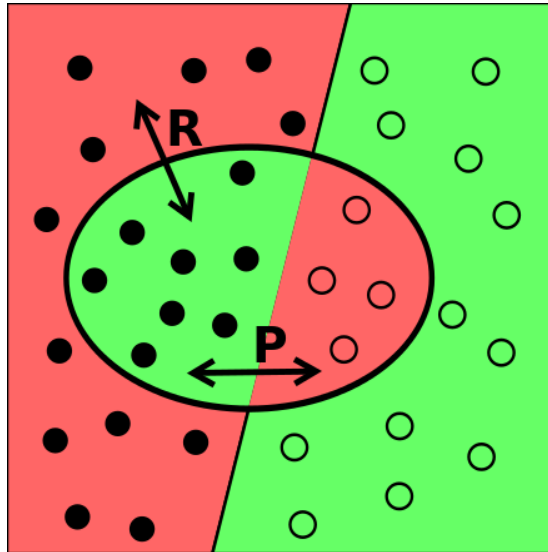


Figura 36: Precision e Recall

In essa gli elementi rilevanti sono tutti quelli che si trovano alla sinistra della retta, mentre gli elementi ritrovati sono quelli appartenenti all'ovale. Le regioni colorate di rosso rappresentano gli errori. In particolare, alla regione rossa di

destra (all'interno dell'ovale) appartengono gli elementi recuperati che, però, non sono rilevanti, ovvero i Falsi Positivi. Invece, alla regione rossa di sinistra (all'esterno dell'ovale) appartengono gli elementi rilevanti che non sono stati recuperati, ovvero i Falsi Negativi. Precision e Recall rappresentano il quoziente dell'area della regione verde con, rispettivamente, l'ovale (freccia orizzontale) e la regione di sinistra (freccia diagonale).

### B.3 F1-Score

Nell'analisi statistica della classificazione binaria, l'F1 score (nota anche come F-score o F-measure, letteralmente "misura F") è una misura dell'accuratezza di un test che tiene in considerazione precision e recall del test. L'F1 viene calcolato tramite la media armonica di precision e recall:

$$F_1 - Score = \frac{2}{\frac{1}{R} + \frac{1}{P}} = 2 * \frac{R * P}{P + R}$$

dove P e R indicano rispettivamente Precision e Recall

La formula generale è:

$$F_\beta - Score = (1 + \beta^2) * \frac{P * R}{(\beta^2 * P) + R}$$

per valori di  $\beta$  reali positivi. Altre istanze di questa formula solitamente utilizzate sono la  $F_2$  (che pone maggiore enfasi sui falsi negativi) e la  $F_{0.5}$  (la quale attenua l'influenza dei falsi negativi).

L'F-score è solitamente usata nel campo del recupero dell'informazione per misurare l'accuratezza delle ricerche o della classificazione dei documenti. Inizialmente l'F1 score era l'unica misura ad essere considerata, ma con la proliferazione in larga scala di motori di ricerca gli obiettivi di prestazione iniziarono a variare, divenendo necessario porre maggiore enfasi su precisione o recupero.

L'F-score è usata anche nel campo dell'apprendimento automatico ed è vastamente impiegata nella letteratura sull'elaborazione del linguaggio naturale.

In particolare, nel nostro caso, l'F1-score è stata utilizzata per entrambi i modelli di apprendimento (sia quello sugli articoli che quello sulle MRI) al fine di stimare l'accuracy sul Test Set utilizzato e determinare il verificarsi del fenomeno dell'overfitting in caso di valori molto elevati.

## C *fastai*

### C.1 CNN con la libreria *fastai*

#### C.1.1 Implementazione

I file DICOM fornitici dagli esperti di dominio sono stati da essi etichettati, e sono state distinte due classi di immagini. In particolare, ci è stato fornito un insieme di 691 immagini, divise in due diverse classi. La prima è la classe GBM che contiene 489 immagini di glioblastomi multiformi, mentre la seconda è la classe MET che contiene 202 immagini di metastasi.

Innanzitutto, i file DICOM etichettati dagli esperti di dominio sono stati suddivisi nei due insiemi di Training e Validation. In particolare, per realizzare la suddivisione è stata sfruttata la tecnica dell'Holdout<sup>5</sup>, per cui il Training Set contiene l'80% dei file, mentre il Validation Set contiene il 20% rimanente.

Ai file DICOM deve essere associato un file **.csv** in cui si specifica l'etichetta associata ad ognuno dei file. Specificando il path in cui è presente il file di etichettamento, il dataset viene caricato utilizzando la classe built-in *DataBlock*:

```
gbm = DataBlock(blocks=(ImageBlock, CategoryBlock),
                 get_x=lambda x: gbm_source/f"{x[0]}",
                 get_y=lambda x: x[1],
                 batch_tfms=aug_transforms(size=300))
```

- *gbm*: variabile che conterrà l'intero *DataBlock*, ovvero l'intero dataset delle risonanze;
- *ImageBlock*: funzione che specifica cosa dovranno contenere i blocchi;
- *gbm\_source*: variabile precedentemente dichiarata contenente il Path del file di labeling delle risonanze.

Caricato il dataset, si generano il Training Set ed il Validation Set che verranno utilizzati per l'addestramento del sistema e la valutazione della sua accuracy. Ciò viene effettuato tramite la funzione *dataloaders* presente all'interno della classe *DataBlock*:

```
dls = gbm.dataloaders(df.values, num_workers=0, bs = int
                      (len(df.values)*0.80), drop_last=True)
```

- *df.values*: carica i valori delle etichette;
- *num\_workers*: indica il numero di sottoprocessi per il caricamento dei dati. Il valore di default 0 indica che viene effettuato dal main process;

---

<sup>5</sup>La tecnica dell'Holdout prevede di suddividere un dataset in modo tale che 2/3 di esso siano usati come Training Set e il restante 1/3 viene usato come Test Set.

- *bs*: indica il dimensionamento dei due Set tramite una suddivisione in **batch**;
- *drop\_last*: elimina eventuali batch generati in eccesso non contenenti elementi del dataset.

È possibile visualizzare il caricamento effettuato sui due set:

```
dls.train.show_batch(max_n=100)
```

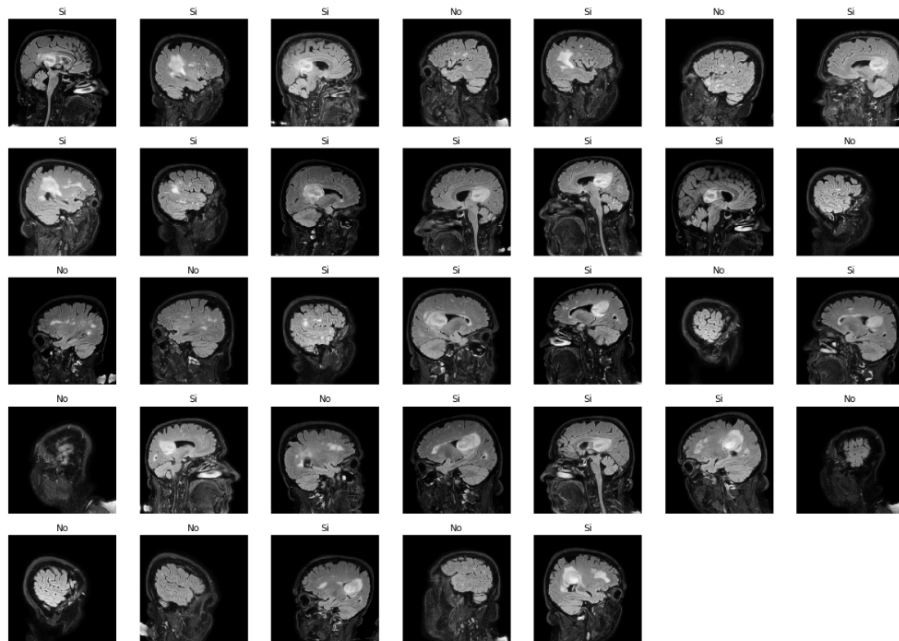


Figura 37: Visualizzazione del Training Set

```
dls.valid.show_batch(max_n=100)
```



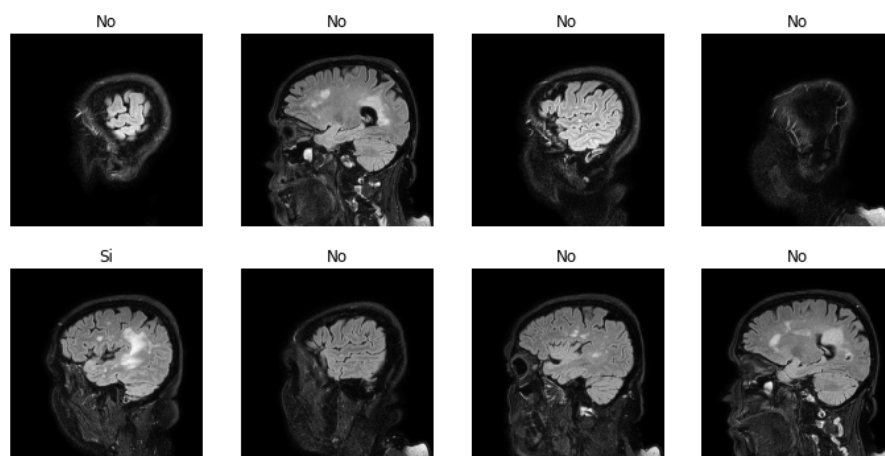


Figura 38: Visualizzazione del Validation Set

### C.1.2 Risultati ottenuti

Generati Training e Validation Sets, è stato effettuato l'addestramento del classificatore, che ha impiegato un tempo pari a 6'03".

Dopodiché, il classificatore ottenuto è stato valutato tramite le metriche di **Accuracy**, **Precision** e **Recall**, ripetendo i cicli di addestramento della CNN fino a quando non raggiungesse un valore di accuracy pari all'83%. Tale valore di soglia è stato scelto in modo tale da evitare il verificarsi del fenomeno dell'overfitting per valori maggiori di accuracy.

Un esempio di addestramento della CNN è osservabile nella figura sottostante:

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.251160	0.729759	0.611111	0.611111	0.611111	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.346079	0.673051	0.666667	0.666667	0.666667	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.293452	0.638695	0.666667	0.666667	0.666667	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.172186	0.592265	0.666667	0.666667	0.666667	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.128820	0.543484	0.666667	0.666667	0.666667	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.085976	0.493860	0.722222	0.722222	0.722222	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.102275	0.446594	0.777778	0.777778	0.777778	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.115888	0.397800	0.777778	0.777778	0.777778	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.095570	0.346179	0.777778	0.777778	0.777778	00:00

epoch	train_loss	valid_loss	accuracy	precision_score	recall_score	time
0	0.075943	0.300927	0.833333	0.833333	0.833333	00:00

Numero di cicli di apprendimento: 34

Figura 39: Esempio di addestramento della CNN

Il risultato della classificazione ottenuto è, quindi, il seguente:

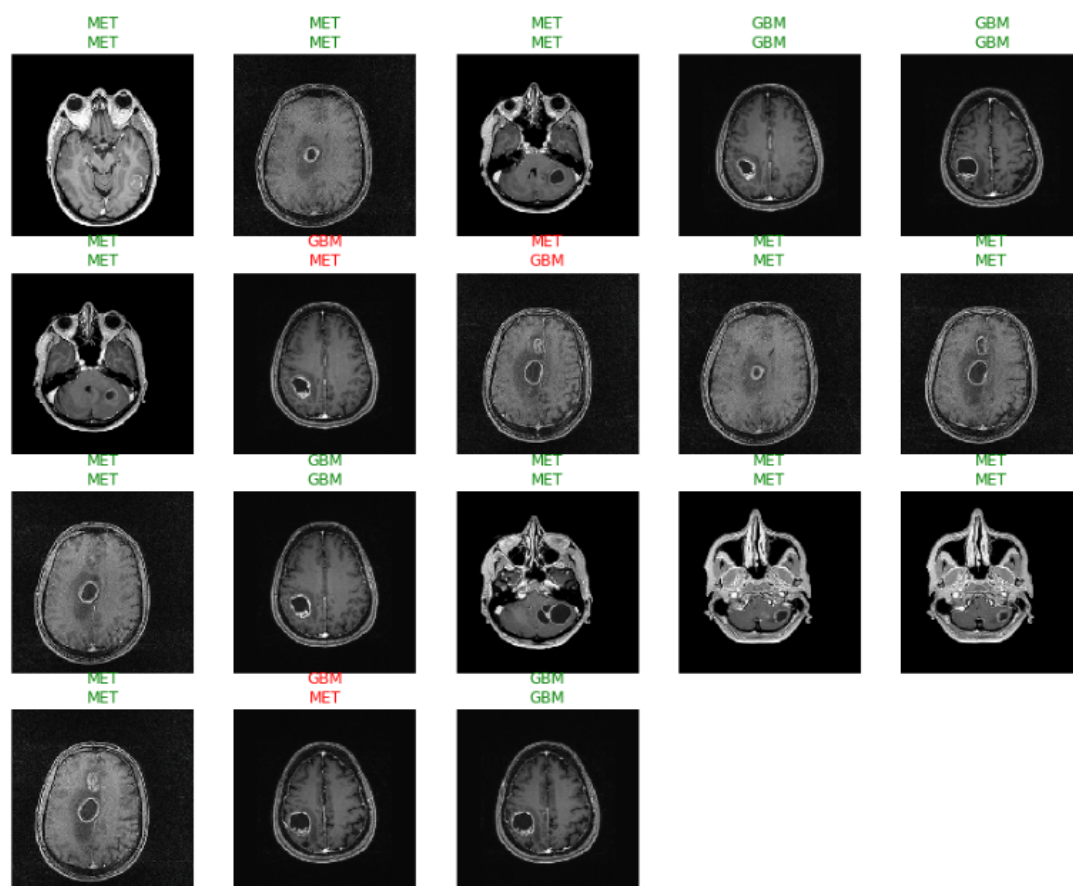


Figura 40: Classificazione effettuata dalla CNN

## D Rimozione testi pediatrici

A seguito di un confronto con gli esperti di dominio, è emerso che dalla query prodotta sono stati recuperati da PubMed articoli di tipo pediatrico che non interessavano per lo studio da effettuare. Dopo sono stati forniti gli id degli articoli da rimuovere, si è prodotto un comando SQL di DELETE per rimuovere i suddetti dal database.

```
DELETE FROM articoligbm  
WHERE pubmedid IN (value1 , value2 , ...)
```

Dove i valori inseriti all'interno della query sono gli id numerici di pubmed forniti dagli esperti.

Al termine dell'operazione, sono stati rimossi 319 articoli dai 1'311 precedentemente presenti.

## Suddivisione del lavoro

- **Guido Maria Secondulfo:** Supervisione generale del progetto, caricamento articoli su Db2, addestramento classificatore di Visual Recognition;
- **Angelo Russo:** Pre-processing sui documenti recuperati, definizione del modello Entità-Relazione su IBM Watson Knowledge Studio;
- **Francesca Lettieri:** Pre-processing sul Test Set delle risonanze, generazione dello script per la valutazione del modello di Visual Recognition e valutazione del classificatore;
- **Attività svolte comunemente:** Definizione query di ricerca, definizione dello script per il recupero delle informazioni degli articoli da PubMed, valutazione del modello di Watson Knowledge Studio, pre-processing sui file DICOM, creazione della documentazione.

## Esperti di dominio

- Dott.ssa Camilla Russo
- Dott.ssa Fernanda Picozzi

## Riferimenti bibliografici

- [1] E-utilities: <https://www.ncbi.nlm.nih.gov/books/NBK25501/>;
- [2] IBM Cloud Db2: <https://www.ibm.com/it-it/cloud/db2-on-cloud>;
- [3] IBM Cloud Foundry:  
<https://www.ibm.com/it-it/cloud/cloud-foundry>;
- [4] GitHub: <https://github.com>;
- [5] MeSH: [https://it.wikipedia.org/wiki/Medical\\_Subject\\_Headings](https://it.wikipedia.org/wiki/Medical_Subject_Headings);
- [6] IBM Watson Machine Learning:  
<https://www.ibm.com/it-it/cloud/machine-learning>;
- [7] IBM Watson Knowledge Studio:  
<https://www.ibm.com/it-it/cloud/watson-knowledge-studio>;
- [8] fastai: <https://www.fast.ai>;
- [9] IBM Cloud API:  
<https://cloud.ibm.com/apidocs/visual-recognition/visual-recognition-v3>;
- [10] Jupyter: <https://jupyter.org>.
- [11] RadiAnt DICOM Viewer: <https://www.radiantviewer.com/it>