UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Rete Neurale per identificazione segnale e rumore in

$$^{17}O(p,\gamma)^{18}F$$

Relatore

Prof. Antonio Caciolli

Correlatore

Dr. Jakub Skowronski

Laureando

Mattia Ponchio

Anno Accademico 2023/2024

Indice

1 Introduzione

1.1 Motivazione Fisica

La reazione $^{17}O + p \rightarrow^{18} F + \gamma$ fa parte del ciclo del carbonio-azoto-ossigeno, che è il meccanismo predominante per la produzione di energia nelle stelle di massa superiore a $1.3M_{\odot}$. In tale processo, isotopi di C, N, O fungono da catalizzatori nella fusione di protoni in nuclei di elio, con produzione di energia, raggi γ e neutrini: il risultato netto dei vari percorsi percorribili nel ciclo può essere riassunto nella seguente maniera:

$$\begin{aligned} 4^1_1H + 2e^- &\rightarrow^4_2 He + 2e^+ + 2e^- + 2\nu_e + 3\gamma + 24.7MeV \\ &\rightarrow^4_2 He + 2\nu_e + 7\gamma + 26.7MeV \end{aligned} \tag{1}$$

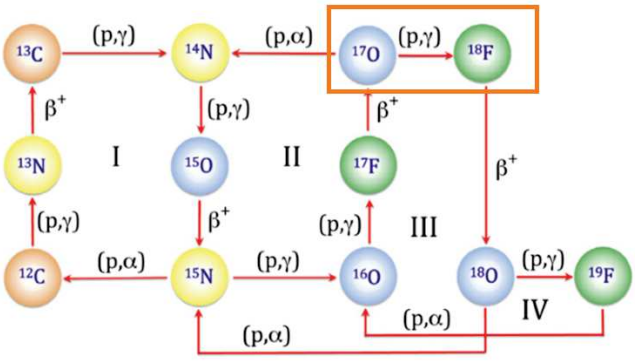

Figura 1: Rappresentazione del ciclo CNO

In particolare, per temperature 20 MK < T < 100 MK la risonanza a 65 keV risulta dominante sulle altre componenti.

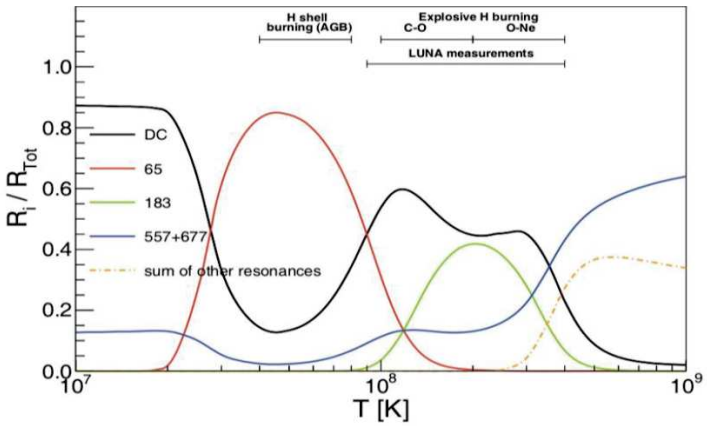

Figura 2: Contributo delle risonanze in funzione della temperatura. Fonte immagine [4].

La forza di questa risonanza è stata determinata attraverso misure indirette e, più recentemente, anche in maniera diretta a LUNA al Gran Sasso (si veda [3]), che grazie alla sua posizione sotterranea riesce a compiere misurazioni che non sarebbero possibili in superficie a causa del fondo cosmico.

Durante lo studio di questa risonanza si è notato che al segnale di $^{17}O(p,\gamma)^{18}F$ (con un Q-valore di $Q_{segnale} = 5607$ keV) si sovrappone un rumore dato da $d(p,\gamma)^3He$ (con un Q-valore di $Q_{rumore} = 5493$ keV), dovuto alla presenza di deuterio contaminante sul target in tantalio utilizzato.

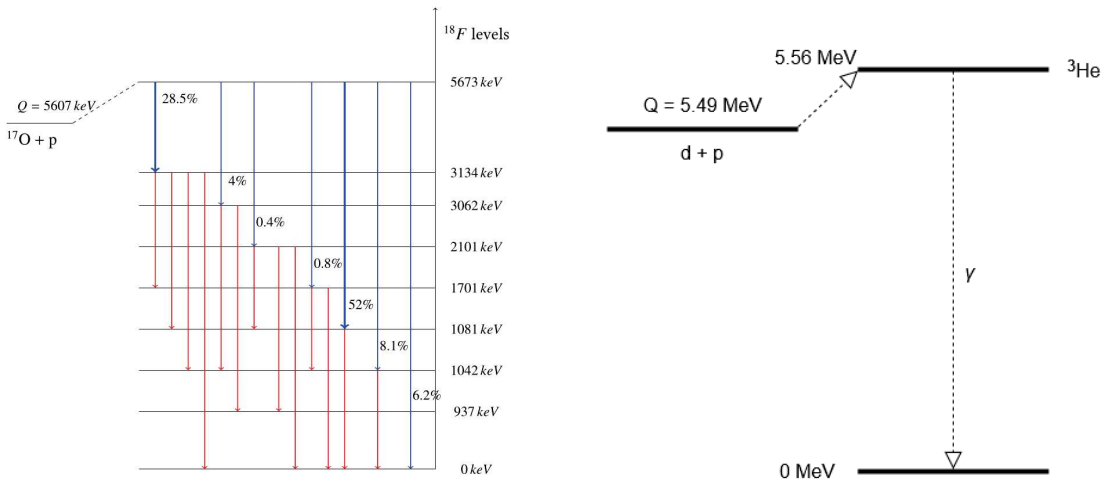

Figura 3: Schema livelli energetici delle due reazioni: per $d(p,\gamma)^3He$ si prevede 1 solo raggio γ, mentre per $^{17}O(p,\gamma)^{18}F$ se ne prevedono, per la maggior parte, 2. Fonte immagine: [1].

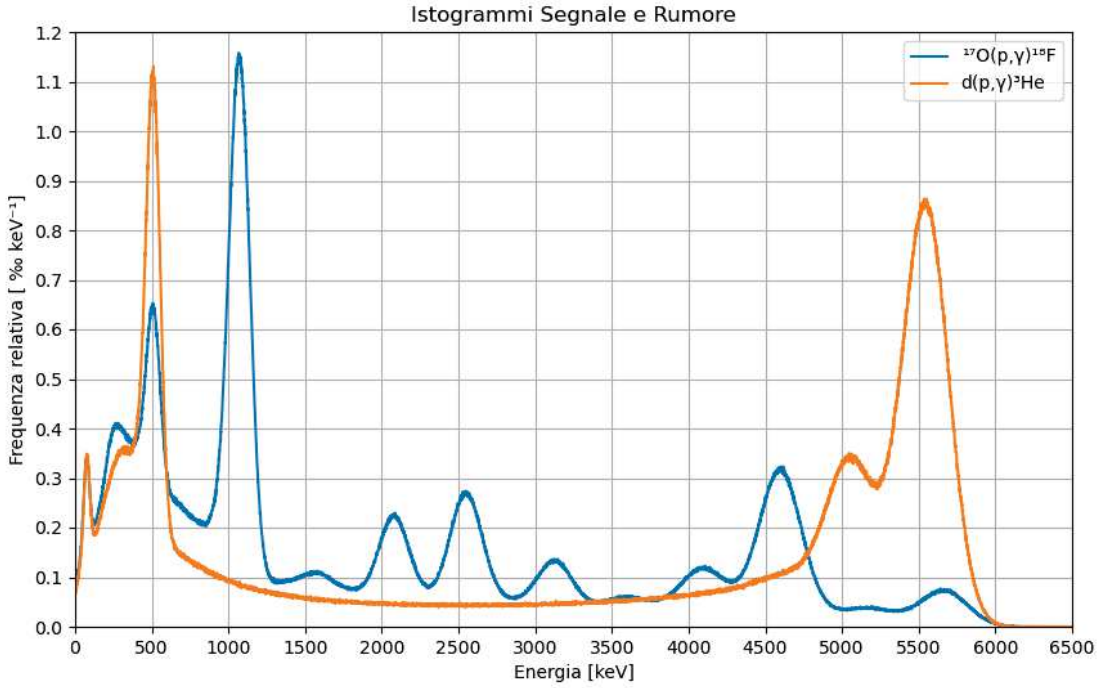

Figura 4: Sovrapposizione istogrammi segnale $^{17}O(p,\gamma)^{18}F$ e rumore $d(p,\gamma)^3He$. Binning = 1 keV.

Grazie alla geometria a quasi 4π e alla segmentazione in 6 cristalli del rivelatore BGO a LUNA è possibile eseguire una analisi delle coincidenze (eventi salvati in list mode), riducendo così il contributo di $d(p,\gamma)^3He$. Infatti, mentre per $^{17}O(p,\gamma)^{18}F$ si aspettano, per la maggior parte, 2 raggi γ per ogni canale di decadimento, per $d(p,\gamma)^3He$ se ne prevede solo 1, con un conseguente effetto sul numero di picchi nello spettro, come si può vedere in figura 4.

Questa Tesi propone un approccio alternativo, non basato su un'analisi evento per evento bensì sullo studio dello spettro in energia in uno dei cristalli per determinare le proporzioni di segnale e rumore presenti nella misura. A tal fine si esplora l'utilizzo di una rete neurale, poiché si intende sfruttare la sua capacità nel pattern recognition per compiere il riconoscimento di segnale e rumore.

1.2 Reti Neurali

Una rete neurale artificiale (Artificial Neural Network o più brevemente NN) prende a modello il funzionamento del cervello umano per creare uno strumento in grado di compiere analisi e predizioni su un set di dati. L'elemento base di questo modello computazionale viene per l'appunto detto neurone (o nodo) e risulta interconnesso a numerosi suoi simili su diversi strati (layers) per formare la rete neurale. Tali strati si possono classificare in tre categorie: il livello di ingresso (input layer), in cui vengono raccolti i dati da analizzare, il livello nascosto (hidden layer), il quale riceve in ingresso i risultati dello strato precedente, compie un'ulteriore analisi su di essi e fornisce i risultati allo strato successivo, e il livello di uscita (output layer), che racchiude le informazioni finali elaborate dalla rete sui dati forniti.

Il funzionamento di una rete neurale dipende fortemente dal tipo di apprendimento scelto, che può essere caratterizzato in *supervisionato*, *non supervisionato*, e *di rinforzo*. Nell'apprendimento supervisionato (supervised learning) vengono forniti assieme ai dati in ingresso anche le corrispettive uscite attese, in maniera tale che la rete possa imparare la relazione tra i due e, dunque, essere anche in grado di fornire predizioni per ingressi la cui uscita sia, a priori, sconosciuta. Nell'apprendimento non supervisionato (unsupervised learning), invece, vengono forniti solamente i dati in ingresso e viene lasciato alla rete il compito di produrre opportuni output, la quale tenta di riconoscere schemi nell'input e raggruppare i dati secondo essi (clustering). Infine, nell'apprendimento di rinforzo (reinforcement learning) la rete cerca di raggiungere un prefissato obiettivo interagendo con l'ambiente, il quale incentiva o disincentiva certe scelte a seconda dello stato attuale e del percorso compiuto.

Per misurare la bontà delle operazioni compiute dalla rete si esamina la cosiddetta funzione di costo (cost function), la cui minimizzazione è alla base dell'apprendimento della rete. Per esempio, nel caso di regressione lineare è possibile scegliere lo scarto quadratico medio come funzione di costo; in generale, dato un modello $g(\theta)$ dipendente dai parametri θ, la funzione di costo $C(\theta, g(\theta))$ permette, tramite la sua minimizzazione in θ, di valutare quanto bene il modello scelto descriva il fenomeno osservato. Tuttavia, capita spesso di non avere accesso alla vera funzione che si desidera minimizzare, utilizzando al suo posto un'approssimazione ottenuta direttamente dai dati e che tale funzione dipenda da molti parametri, ritrovandosi così con una funzione in uno spazio ad alta dimensione con molti minimi locali.

Il metodo basilare con cui minimizzare $C(\theta, g(\theta))$, spesso chiamata anche energia $E(\theta)$ in analogia a sistemi fisici, è il metodo della discesa del gradiente (gradient descent, *GD*), in cui si inizializzano i parametri a dei valori arbitrari θ_0 e si aggiornano secondo le formule:

$$\begin{aligned} v_t &= \eta_t \nabla_\theta E(\theta_t) \\ \theta_{t+1} &= \theta_t - v_t \end{aligned} \tag{2}$$

dove si è introdotto il tasso di apprendimento η_t (learning rate), che controlla la lunghezza degli spostamenti lungo la direzione data dal gradiente. Tale grandezza fa parte degli *iperparametri* della rete, cioè di quei parametri che costituiscono la rete e ne determinano l'efficienza. La scelta del tasso da utilizzare è importante perché con un η_t troppo piccolo, seppur si arrivi al minimo locale, si paga un alto costo computazionale, mentre nel caso di un η_t troppo grande il metodo diventa instabile, oscillando attorno al minimo o addirittura allontanandosi da esso.

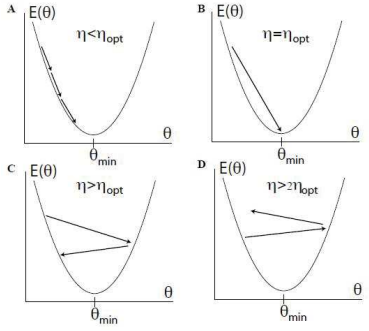

Figura 5: Esempio dell'effetto della scelta del tasso di apprendimento nel caso di un potenziale quadratico unidimensionale, con $\eta_{opt} = [\partial^2_\theta E(\theta)]^{-1}$. Fonte immagine [2].

Questo metodo, però, presenta molte carenze: innanzitutto, come si è visto, è molto sensibile alla scelta del tasso di apprendimento e, allo stesso tempo, anche al dato iniziale, in quanto esso determinerà il minimo locale individuato. Inoltre comporta un alto costo computazionale quando si hanno molti dati, in quanto richiede il calcolo di un gradiente per ogni punto del dataset, e impiega tempi che crescono esponenzialmente per fuggire dai punti di sella.

L'algoritmo viene, dunque, migliorato tramite la discesa stocastica del gradiente (stochastic gradient descent*, SGD*), nel quale si formano dei sottogruppi chiamati *minibatches* e si approssima il calcolo del gradiente su ogni punto al calcolo del gradiente su un singolo minibatch. Un ciclo su tutti questi ultimi viene chiamato epoca (epoch):

$$\begin{aligned} &\nabla_\theta E^{MB}(\theta) = \sum_{i \in B_k} \nabla_\theta e_i(x_i, \theta) \\ &v_t = \eta_t \nabla_\theta E^{MB}(\theta_t) \\ &\theta_{t+1} = \theta_t - v_t \end{aligned} \tag{3}$$

I vantaggi di questo approccio stocastico includono un incremento nella probabilità di fuggire a punti di minimo locale isolati e un minor tempo di esecuzione grazie all'approssimazione in sottogruppi.

Un'ulteriore raffinamento di questo metodo prevede l'introduzione di un termine di "momento" (o di inerzia) γ, che mantiene l'informazione della direzione dello spostamento nello spazio delle fasi effettuato nel passo precedente:

$$\begin{aligned} &v_t = \gamma v_{t-1} + \eta_t \nabla_\theta E^{MB}(\theta_t) \\ &\theta_{t+1} = \theta_t - v_t \end{aligned} \tag{4}$$

In questa maniera v_t risulta la media mobile dei gradienti nel tempo caratteristico $(1-\gamma)^{-1}$. I vantaggi di questo approccio rendono possibile per l'algoritmo di guadagnare velocità nelle direzioni in cui si hanno gradienti consistentemente piccoli e di sopprimere le oscillazioni nelle direzioni con grande curvatura.

Infine, indipendentemente dall'utilizzo del momento, è comunque necessario specificare la regola di aggiornamento per il il tasso di apprendimento, idealmente in maniera tale per cui si compiano grandi passi nelle direzioni "pianeggianti" con poca pendenza e piccoli passi nelle direzioni in cui si hanno, invece, elevate pendenze. Un esempio di questo tipo di algoritmo, utilizzato anche in questa trattazione, è ADAM, in cui si tiene conto non solo del primo momento di aspettazione del gradiente $m_t = \mathbb{E}[g_t]$ ma anche del secondo $s_t = \mathbb{E}[g_t^2]$ per ottenere i comportamenti desiderati:

$$\begin{aligned} &g_t = \nabla_\theta E(\theta) \\ &m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ &s_t = \beta_2 s_{t-1} + (1-\beta_2) g_t^2 \\ &\hat{m}_t = \frac{m_t}{1-(\beta_1)^t} \\ &\hat{s}_t = \frac{s_t}{1-(\beta_2)^t} \\ &\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{s}_t} + \epsilon} \end{aligned} \tag{5}$$

$\hat{m}_t$ e $\hat{s}_t$ sono medie mobili del primo e secondo momento di aspettazione del gradiente. β_1 e β_2 sono i tempi tipici per le medie mobili e sono presi solitamente essere rispettivamente 0.9 e 0.99. η_t è un tasso di apprendimento di riferimento, usualmente scelto come 10^{-3}, mentre $\eta \sim 10^{-8}$ è una piccola regolarizzazione per evitare divergenze (formule tratte da [2]).

Un ulteriore vantaggio di ADAM è che la discriminazione di gradienti in "grandi" e "piccoli" si può trattare in termini della scala naturale della deviazione standard $\sigma_t = \sqrt{\hat{s}_t - (\hat{m}_t)^2}$: infatti si può scrivere la lunghezza dello spostamento per un singolo parametro θ_t come

$$\Delta\theta_{t+1} = -\eta_t \frac{\hat{m}_t}{\sqrt{\sigma_t^2 + \hat{m}_t^2 + \epsilon}} \tag{6}$$

Si può, dunque, osservare come nel caso di varianza piccola (e $\hat{m}_t \gg \epsilon$) si abbia $\Delta\theta_{t+1} \to -\eta_t$, limitando così la lunghezza massima dei passi nelle direzioni ripide, mentre nel caso ad alta varianza $\sigma^2 \gg \hat{m}_t^2$ risulti $\Delta\theta_{t+1} \to -\eta_t \frac{\hat{m}_t}{\sigma_t}$, adattando in questa maniera il tasso di apprendimento in proporzione al valore medio del primo momento di aspettazione del gradiente espresso in unità di deviazione standard.

2 Raccolta Dati

Gli istogrammi di energia sono stati creati tramite simulazione con Geant4, in cui si è tenuto conto delle caratteristiche misurate sperimentalmente per la modellizzazione di ogni cristallo. Il background intrinseco al rivelatore, invece, non è stato incluso in questa trattazione per semplificazione. Inoltre, si è effettuato un taglio sulle energie analizzate, scegliendo le sole simulazioni in cui l'energia totale depositata sui cristalli fosse compresa tra 5000 keV e 6500 keV, che risulta comprendere la regione di interesse.

Per ogni cristallo sono stati generati 1000 istogrammi, riempiti con 1000000 di elementi ciascuno, con un binning di 1 keV. Per ogni istogramma la frazione di segnale $^{17}O(p,\gamma)^{18}F$ viene scelta tramite generazione di un numero casuale $s \in (0,1)$ in precisione singola con una distribuzione uniforme. L'istogramma viene, dunque, riempito per una tale frazione s dei conteggi totali con gli elementi relativi allo spettro di $^{17}O(p,\gamma)^{18}F$, sempre selezionati utilizzando una distribuzione uniforme per non introdurre alcun bias, e lo stesso viene fatto per la rimanente porzione dell'istogramma con il rumore $d(p,\gamma)^3He$.

2.1 Architettura della rete

La struttura della rete utilizzata viene detta *Encoder*, e risulta organizzata nella seguente maniera:

- *Strato di ingresso*: contiene i conteggi degli istogrammi con le relative proporzioni di segnale e rumore. Le frequenze vengono inoltre convertite da assolute in relative per svincolare l'analisi dal numero di conteggi totali ed espresse in ‰ in quanto, empiricamente, si verifica che la rete converge all'individuazione delle caratteristiche dei dati più velocemente quando questi hanno valori prossimi all'ordine di grandezza dell'unità.
- *Strati nascosti*: vi sono 3 livelli nascosti, che agiscono tutti nella stessa maniera sui dati per trovare la relazione che lega l'ingresso x all'uscita y. Le due operazioni svolte da questi strati sono una trasformazione lineare $y = xA^T + b$ (metodo *Linear*) seguita da una funzione di unità lineare rettificata $y = max(0,x)$ (metodo ReLU).

  Tali operazioni, oltre a migliorare la stima della funzione che lega *input* e *output*, permettono anche di ridurre il numero di caratteristiche necessarie alla rete per riconoscere ed analizzare diversi ingressi con ogni strato consecutivo, cioè all'aumentare della profondità della rete è possibile concentrare le informazioni peculiari dei dati in ingresso in poche variabili. Il numero di neuroni viene perciò modificato secondo lo schema:

  $input \to out1 \to out2 \to output$

  dove, nell'analisi eseguita, si è scelto di porre $input = 10000$ e $output = 2$ per fare corrispondere l'ingresso al numero di bin per istogramma e l'uscita ai due valori di proporzione segnale-rumore.
- *Strato di uscita*: contiene le due grandezze caratteristiche estratte dalla rete sui dati forniti, corrispondenti alle proporzioni di segnale e rumore per ogni istogramma.

Il dataset viene diviso in maniera casuale per il 70% nell'insieme di allenamento, per il 10% nell'insieme di test e il restante 20% nell'insieme di validazione, con *minibatches* di 64 elementi ed un totale di 100 epoche.

Come funzione di costo si è utilizzato l'errore quadratico medio (MSE) tra le previsione delle proporzioni segnale-rumore elaborate dalla rete e quelle reali su tutti gli istogrammi, mentre per stimare l'accuratezza del metodo si è utilizzato il coefficiente di determinazione R^2.

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^{n}(y_{i,rete} - y_i)^2}{n} \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_{i,rete} - y_i)^2}{\sum_{i=1}^{n}(y_{i,rete} - \overline{y}_{rete})^2} \end{aligned} \tag{7}$$

R^2 è un indicatore della bontà con cui il modello scelto descrive i dati esaminati: ad un $R^2 = 1$ corrisponde un modello che perfettamente descrive il fenomeno osservato. È perciò interpretabile come la frazione della variabilità della variabile dipendente sia spiegabile tramite la variabile indipendente scelta ed il modello utilizzato.

In questo caso si utilizza il coefficiente di determinazione per valutare quanto bene le 2 caratteristiche estratte dalla rete tramite l'analisi delle frequenze spieghino le frazioni di segnale e rumore reali, permettendo così un confronto fra le porzioni previste dalla rete e quelle reali.

3 Analisi dati

Modificando gli iperparametri *out1* e *out2*, ovvero il numero di neuroni in uscita dal primo e secondo strato nascosto, si influenza la velocità con cui la rete acquisisce informazioni e converge alle caratteristiche desiderate di proporzione segnale e rumore. Si è, dunque, proceduto ad ottimizzare la rete modificando questi valori, tenendo innanzitutto fisso out2 ad un basso valore arbitrario (ciò sarà motivato in seguito) e variando out1 fino a trovarne un valore ottimale, per poi studiare anche l'effetto della variazione di out2 su quest'ultimo. Poiché si desidera diminuire il numero di neuroni all'incrementare della profondità della rete, si è scelto come valore iniziale out2=128, in maniera da poter esplorare per out1 la regione di valori compresa tra 10000 e 128.

3.1 Ottimizzazione della rete

L'ottimizzazione della rete è stata effettuata studiando il comportamento della funzione di costo e dell'accuratezza nell'addestramento lungo le varie epoche al variare del numero di neuroni negli strati nascosti.

Per ogni coppia $(out1, out2)$ si è allenata la rete con tali iperparametri per 10 volte, fornendo come valori rappresentativi dei campioni i valori medi delle epoche per cui la funzione di costo o l'accuratezza soddisfano la soglia indicata. L'incertezza associata è l'errore della media, ottenuta come $\frac{\sigma}{\sqrt{10}}$, con σ deviazione standard. Si espone, innanzitutto, l'andamento al variare dell'iperparametro *out1*.

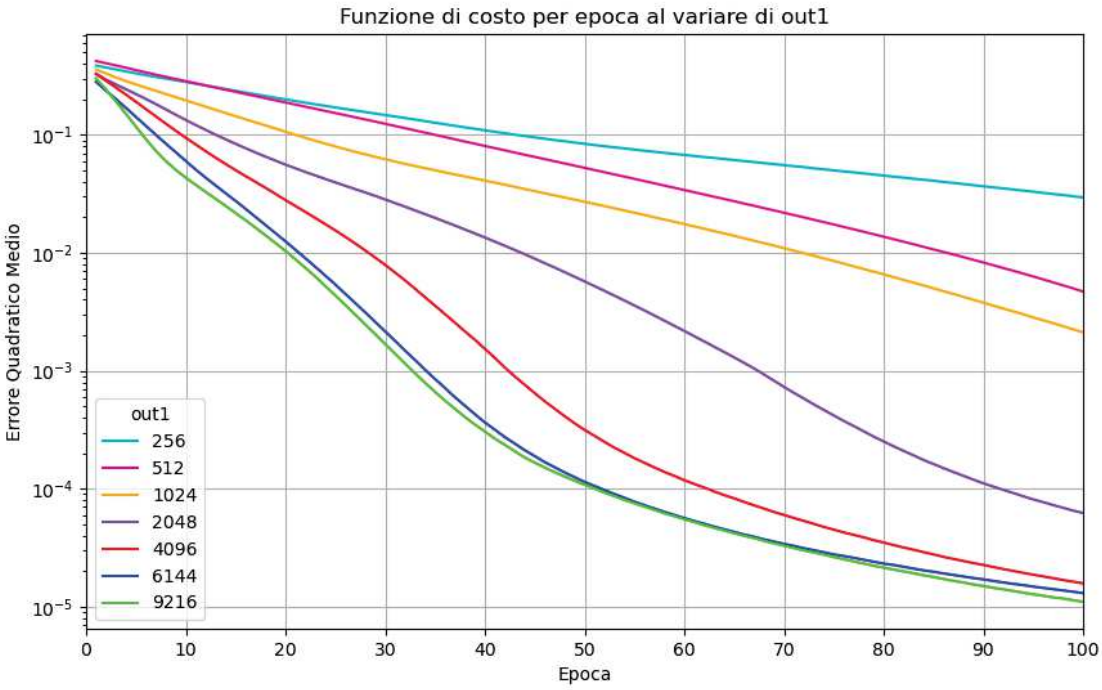

Figura 6: Funzione di costo nell'insieme di allenamento al variare di *out1* in funzione dell'epoca. Asse y in scala logaritmica.

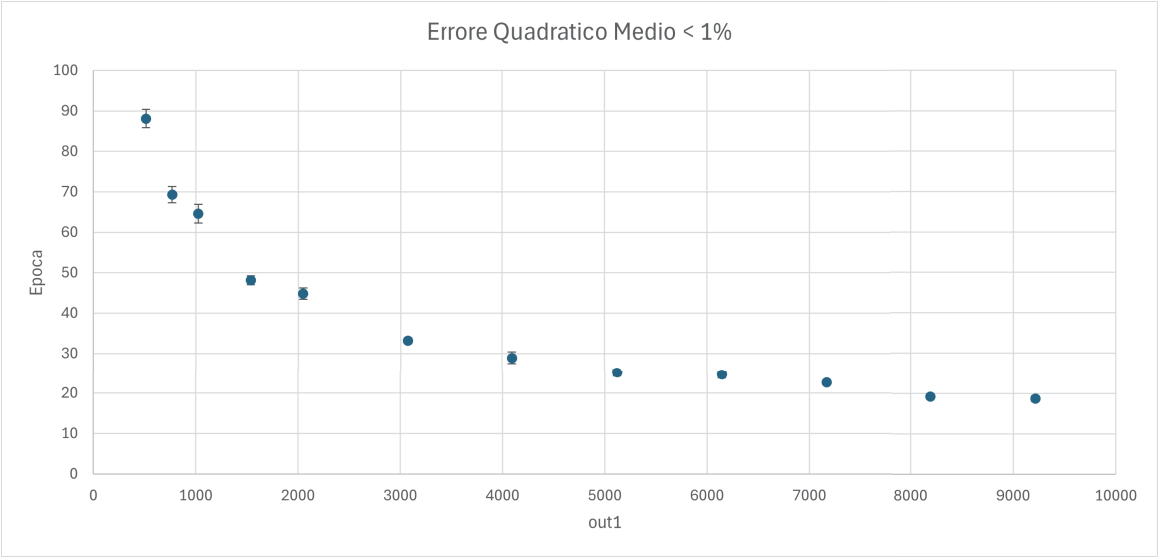

Figura 7: Prima epoca per cui l'errore quadratico medio sia mediamente inferiore alla soglia scelta in funzione di *out1*.

	Epoca media per cui			
out1	MSE $< 20\%$	MSE $< 10\%$	MSE $< 5\%$	MSE $< 1\%$
256	22±2	45±2	72±2	/
512	14±1	30±2	48±2	88±2
768	12.7±0.7	26±1	39±1	69±2
1024	10.9±0.8	21±1	34±1	69±2
1536	8.2±0.6	16.5±0.7	25.7±0.8	48±1
2048	7.8±0.4	15.3±0.6	24.2±0.8	45±1
3072	6.3±0.3	11.5±0.5	17.6±0.4	33.2±0.7
4096	5.1±0.4	9.8±0.6	14.9±0.9	28.9±1.4
5120	4.3±0.2	8.6±0.3	12.9±0.5	25.1±0.8
6144	4.5±0.2	8.6±0.3	12.8±0.4	24.6±0.9
7168	4.5±0.2	8±0.2	11.8±0.3	22.6±0.6
8192	3.8±0.2	7±0.2	10.4±0.2	19.2±0.6
9216	3.8±0.2	6.8±0.2	10.1±0.4	18.7±0.7

Tabella 1: Prima epoca per cui il Mean Squared Error risulti mediamente inferiore alle varie soglie. Grafico in figura 7 rappresentativo dell'andamento per tutte le soglie.

Come si evince dal grafico 7, la rete tende a convergere più velocemente al crescere del numero di neuroni in uscita dal primo strato nascosto: questo comportamento si ripete indipendentemente dalla soglia per l'errore quadratico medio scelta. Tale andamento è anche confermato dall'accuratezza, per cui si riportano analoghi grafici in figure 8 e 9.

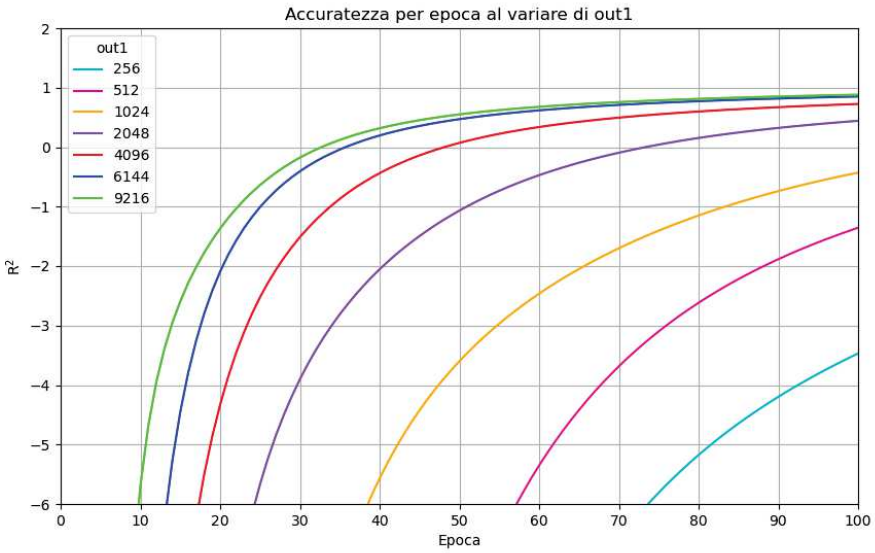

Figura 8: Accuratezza nell'insieme di allenamento al variare di *out1* in funzione dell'epoca.

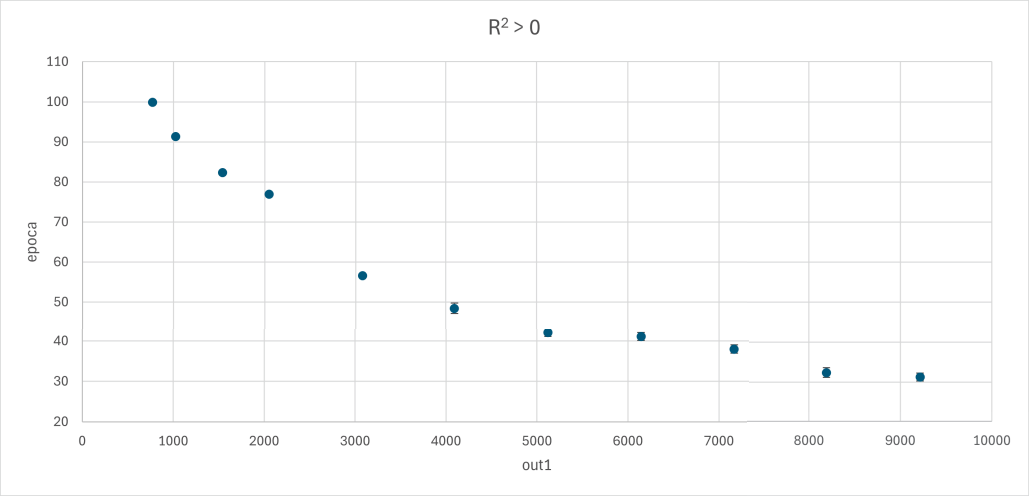

Figura 9: Prima epoca per cui il coefficiente di determinazione sia superiore alla soglia scelta in funzione di *out1*.

	Epoca media per cui		
out1	$R^2 > 0$	$R^2 > 0.50$	$R^2 > 0.80$
256	/	/	/
512	/	/	/
768	99.9±0.1	/	/
1024	91.3±0.4	/	/
1536	82.31±0.4	/	/
2048	76.9±0.1	99±1	/
3072	56.6±0.4	83.3±1.9	/
4096	48.4±1.3	71.7±3.9	98±2
5120	42±1	61±2	96±2
6144	41±1	60±2	96±2
7168	38±1	56±1	94±2
8192	32±1	48±1	81±2
9216	31±1	46±2	77±3

Tabella 2: Prima epoca per cui il coefficiente di determinazione R^2 risulti mediamente superiore alle varie soglie. Grafico in figura 9 rappresentativo dell'andamento per tutte le soglie.

Dunque, poiché la velocità di convergenza della rete in termini di epoche risulta aumentare all'incrementare del numero di neuroni in uscita dal primo strato nascosto, si è fissato *out1 = 9216*. Per studiare l'ulteriore effetto sulla rapidità di convergenza al variare del numero di neuroni in uscita dal secondo strato nascosto, si è proceduto con un'analisi analoga per l'iperparametro *out2*, tenendo conto della restrizione $2 < \text{out2} < 9216$.

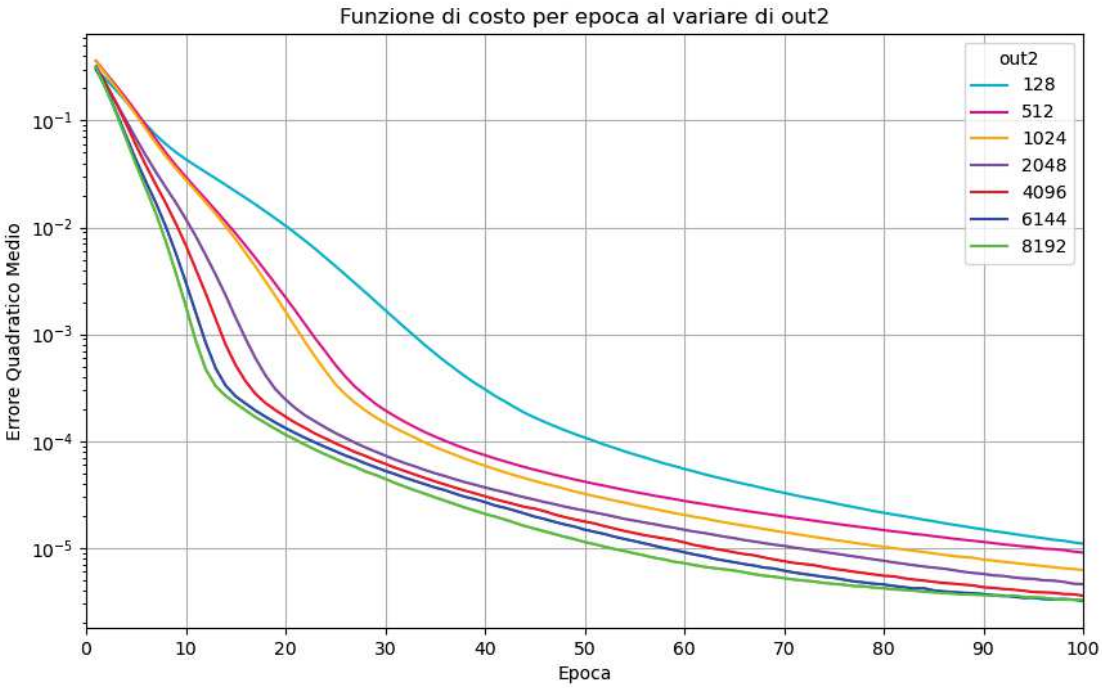

Figura 10: Funzione di costo nell'insieme di allenamento al variare di *out2* in funzione dell'epoca. Asse y in scala logaritmica.

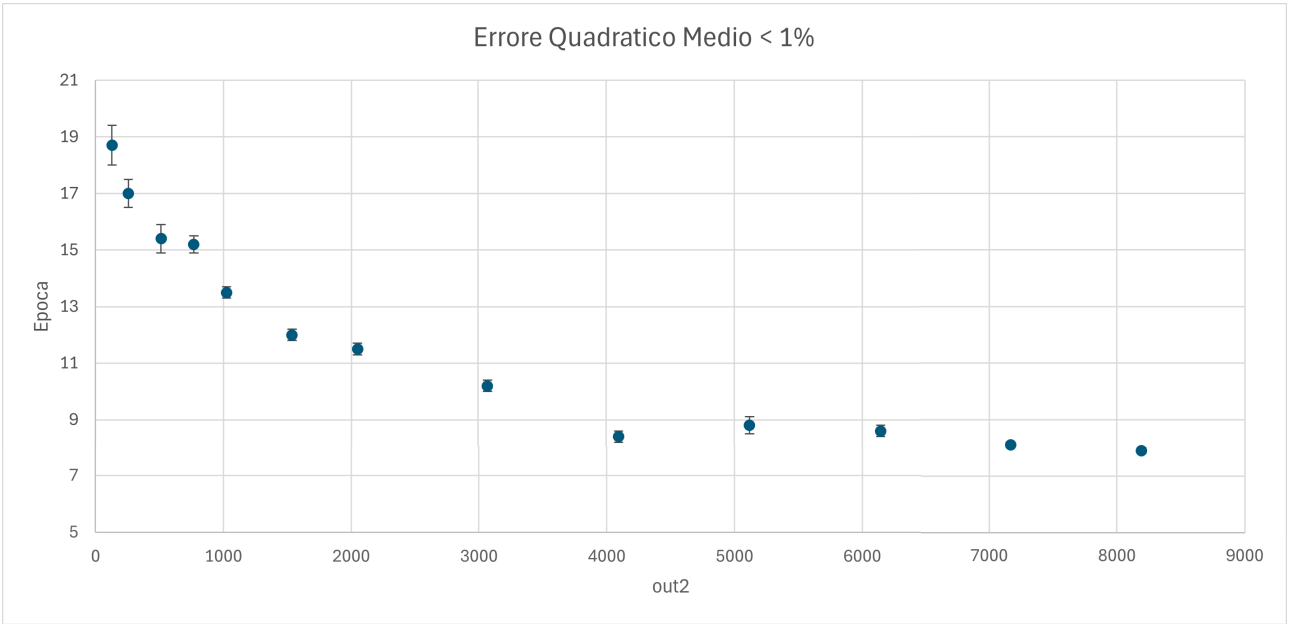

Figura 11: Prima epoca per cui l'errore quadratico medio sia inferiore alla soglia scelta in funzione di *out2*.

	Epoca media per cui			
out2	MSE $< 20\%$	MSE $< 10\%$	MSE $< 5\%$	MSE $< 1\%$
128	3.8±0.2	6.8±0.2	10.1±0.4	18.7±0.7
256	3.4±0.4	6.4±0.4	9.0±0.3	17±0.5
512	3.6±0.2	6.4±0.2	8.6±0.2	15.4±0.5
768	3.5±0.5	5.7±0.4	8.5±0.4	15.2±0.3
1024	3.2±0.1	5.5±0.2	7.6±0.2	13.5±0.2
1536	3.1±0.1	5.1±0.1	7.2±0.3	12.0±0.2
2048	3.0±0.1	4.8±0.2	6.5±0.2	11.5±0.2
3072	2.9±0.1	4.4±0.2	6.1±0.1	10.2±0.2
4096	2.6±0.2	4.4±0.2	5.5±0.2	8.4±0.2
5120	3.0±0.4	4.4±0.2	5.4±0.2	8.8±0.3
6144	2.5±0.6	4.2±0.1	5.4±0.3	8.6±0.2
7168	2.4±0.2	4.1±0.1	5.1±0.1	8.1±0.1
8192	2.4±0.2	3.9±0.1	4.9±0.1	7.9±0.1

Tabella 3: Prima epoca per cui il *Mean Squared Error* risulti mediamente inferiore alle varie soglie. Grafico in figura 11 rappresentativo dell'andamento per tutte le soglie.

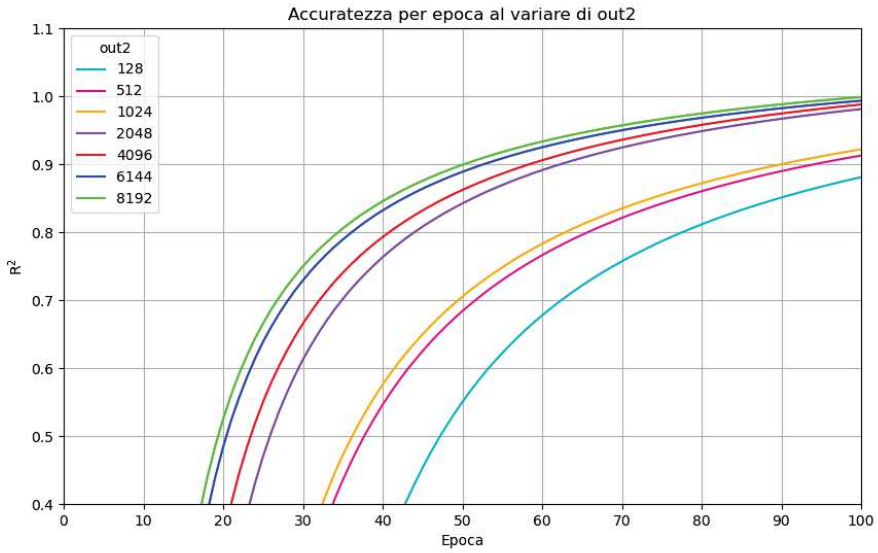

Figura 12: Accuratezza nell'insieme di allenamento al variare di out2 in funzione dell'epoca.

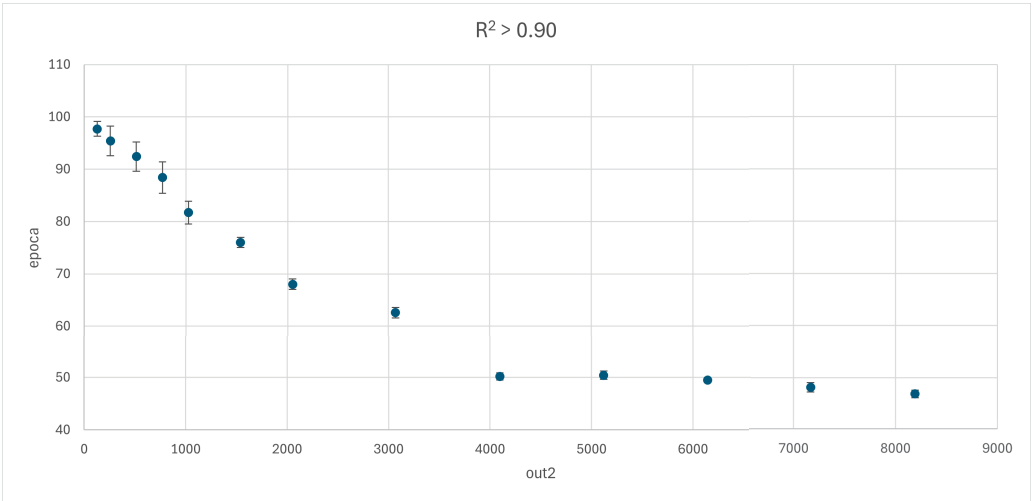

Figura 13: Prima epoca per cui il coefficiente di determinazione sia superiore alla soglia scelta in funzione di out2.

	Epoca media per cui				
out2	$R^2 > 0.50$	$R^2 > 0.80$	$R^2 > 0.90$	$R2 > 0.95$	$R^2 > 0.99$
128	46±2	77±3	98±1	/	/
256	42±2	71±4	95±3	/	/
512	38±1	66±2	92±3	/	/
768	37.6±0.6	66±3	88±3	/	/
1024	33.7±0.7	58±1	82±2	98.8±0.8	/
1536	30.4±0.4	53.0±0.9	76±1	98±2	/
2048	27.9±0.6	47.9±0.9	68±1	88±2	/
3072	25.6±0.4	44.2±0.8	63±1	82±1	/
4096	20.6±0.2	35.4±0.5	50.2±0.7	65.4±0.9	99±1
5120	20.2±0.2	36±1	50.4±0.8	65±1	93±1
6144	20.2±0.3	34±1	49.5±0.5	64±1	87±1
7168	20.0±0.3	33.3±0.6	48.2±0.9	63±1	82±2
8192	19.0±0.3	33.0±0.4	46.9±0.7	61±1	81±2

Tabella 4: Prima epoca per cui il coefficiente di determinazione R^2 risulti mediamente superiore alle varie soglie. Grafico in figura 13 rappresentativo dell'andamento per tutte le soglie.

Per quanto riguarda la convergenza della funzione di costo e dell'accuratezza, la rete neurale sembra prediligere nuovamente un elevato numero di neuroni nello strato nascosto considerato: con out1 fissato, un aumento di ou2 produce un miglioramento nella prestazione della rete. Una possibile spiegazione di questo fenomeno è che la rete abbia una prestazione tanto migliore quanto sia superiore il numero di neuroni totali negli strati nascosti, e dunque dipenda principalmente dalla somma $out1 + out2$ piuttosto che dai due iperparametri considerati singolarmente. Per verificare tale ipotesi si è già mostrato, innanzitutto, come, fissando uno qualsiasi tra out1 e out2 e incrementando l'altro, la velocità di convergenza della rete aumenti. Si procede, perciò, offrendo un confronto per la prestazione dell'Encoder tramite valutazione della funzione di costo e dell'accuratezza nel campione di addestramento a seconda del numero totale di neuroni negli strati nascosti.

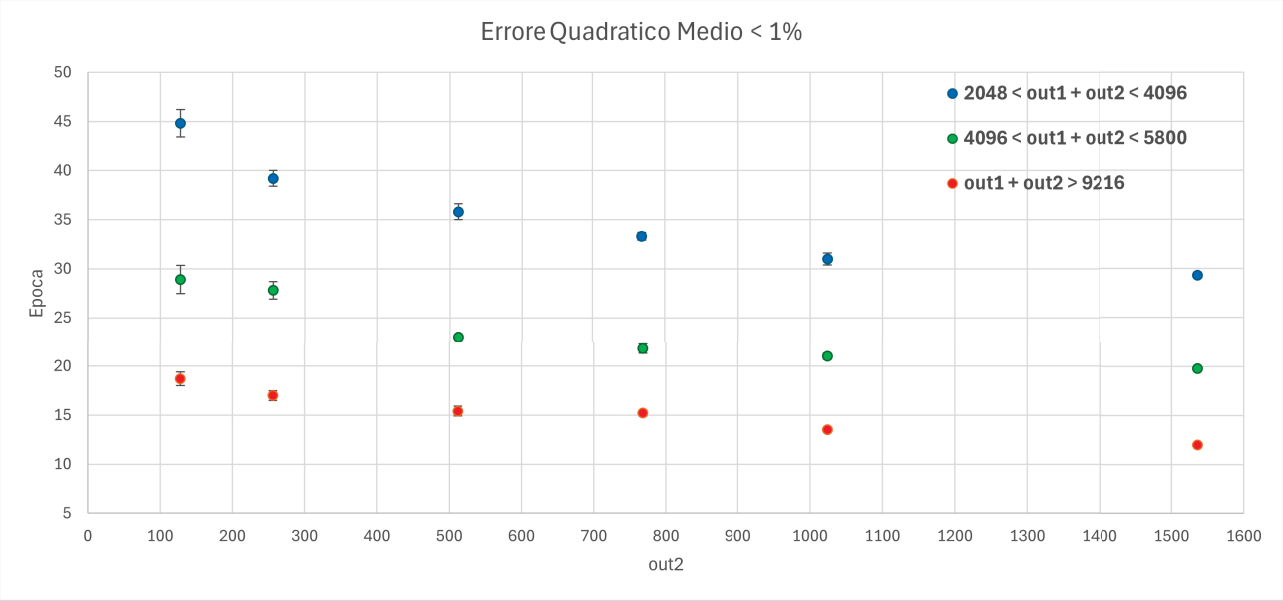

Figura 14: In blu out1=2048, in verde out1=4096, in rosso out1=9216. Si nota come, fissato out2, aumentare out1 migliori la prestazione della rete. Inoltre per i tre set di dati si mette in evidenza come all'aumentare del numero totale di neuroni negli strati nascosti out1+out2 la prestazione della rete migliori. Rispetto ai punti in blu, dunque, quelli verdi corrispondono a 2048 neuroni totali in più e quelli rossi a 7168 in più.

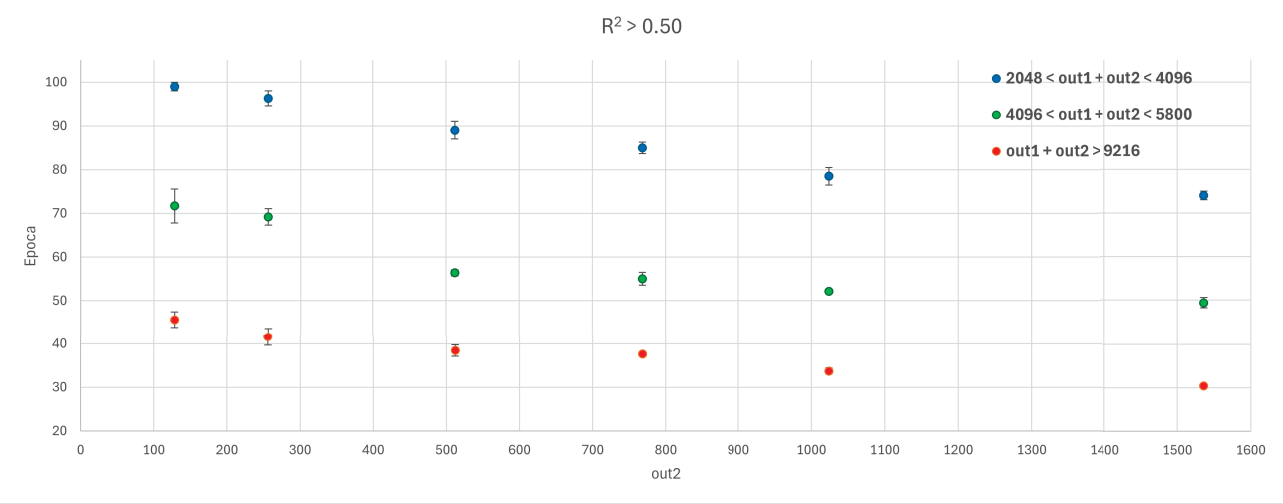

Figura 15: Come in figura 14: in blu out1=2048, in verde out1=4096, in rosso out1=9216.

out1 = 2048	
out2	MSE < 1%
128	45±1
256	39.2±0.8
512	35.8±0.8
768	33.3±0.4
1024	31.0±0.6
1536	29.3±0.3

out1 = 4096	
out2	MSE < 1%
128	29±1
256	27.8±0.9
512	23.0±0.2
768	21.8±0.5
1024	21.0±0.3
1536	19.8±0.2

out1 = 9216	
out2	MSE < 1%
128	18.7±0.7
256	17.0±0.5
512	15.4±0.5
768	15.2±0.3
1024	13.5±0.2
1536	12.0±0.2

Tabella 5: Tabelle relative agli errori quadratici medi in figura 14 a seconda di out1: da sinistra verso destra si hanno rispettivamente i punti blu, verdi e rossi.

out1 = 2048	
out2	$R^2 > 0.50$
128	99±1
256	96±2
512	89±2
768	85±1
1024	79±2
1536	74±1

out1 = 4096	
out2	$R^2 > 0.50$
128	72±4
256	69±2
512	56±1
768	55±1
1024	52±1
1536	49±1

out1 = 9216	
out2	$R^2 > 0.50$
128	46±2
256	42±2
512	38±1
768	37±1
1024	34±1
1536	30±1

Tabella 6: Tabelle relative ai coefficienti di determinazione in figura 15 a seconda di out1: da sinistra verso destra si hanno rispettivamente i punti blu, verdi e rossi.

Si può, perciò, affermare che il fattore principale nella prestazione della rete sia il numero di neuroni totali negli strati nascosti, piuttosto che la loro distribuzione sui singoli strati interni. In particolare, la velocità di convergenza della rete migliora all'aumentare del numero totale di features negli strati interni e conviene, per questa ragione, scegliere un out1 elevato in maniera da poter fare lo stesso per out2.

Tuttavia, incrementare il numero di neuroni degli strati interni alla rete aumenta il tempo computazionale necessario per il suo allenamento e, in questo caso, si ha necessità di diminuire progressivamente la dimensione della rete fino a ottenere uno strato di uscita con 2 neuroni, rappresentativi delle predizioni compiute dalla rete sulle proporzioni di segnale e rumore per lo spettro analizzato.

Dunque, considerati i comportamenti proposti per funzione di costo e accuratezza, la rete risulta ottimizzata utilizzando gli iperparametri:

$$\begin{aligned} out1 &= 9216 \\ out2 &= 8192 \end{aligned} \tag{8}$$

3.2 Analisi dei risultati

Le considerazioni sulla bontà della prestazione della rete esposte sono principalmente indicatori sulla sua capacità di fittare i dati con cui è stata allenata. Tale processo è, inoltre, ottenuto come una mediazione su tutti gli istogrammi utilizzati e sulle minibatches create e non fornisce, per esempio, informazioni sulla bontà delle predizioni in funzione delle frazioni di segnale e rumore presenti o sulla capacità delle rete di discernere fra diversi istogrammi. Si intende, per questo motivo, fornire degli indicatori sulla capacità delle rete di compiere tali compiti.

A tale scopo sono stati, dunque, forniti alla rete degli istogrammi misti con frazione di segnale tra 1% e 99%, e si è confrontata la predizione della rete su di esse con il loro vero valore in funzione della frazione di rumore $d(p,\gamma)^3He$ (figura 17). Si riportano di seguito alcuni di questi istogrammi utilizzati, considerati significativi per intuire la variazione dello spettro al variare della presenza di segnale $^{17}O(p,\gamma)^{18}F$.

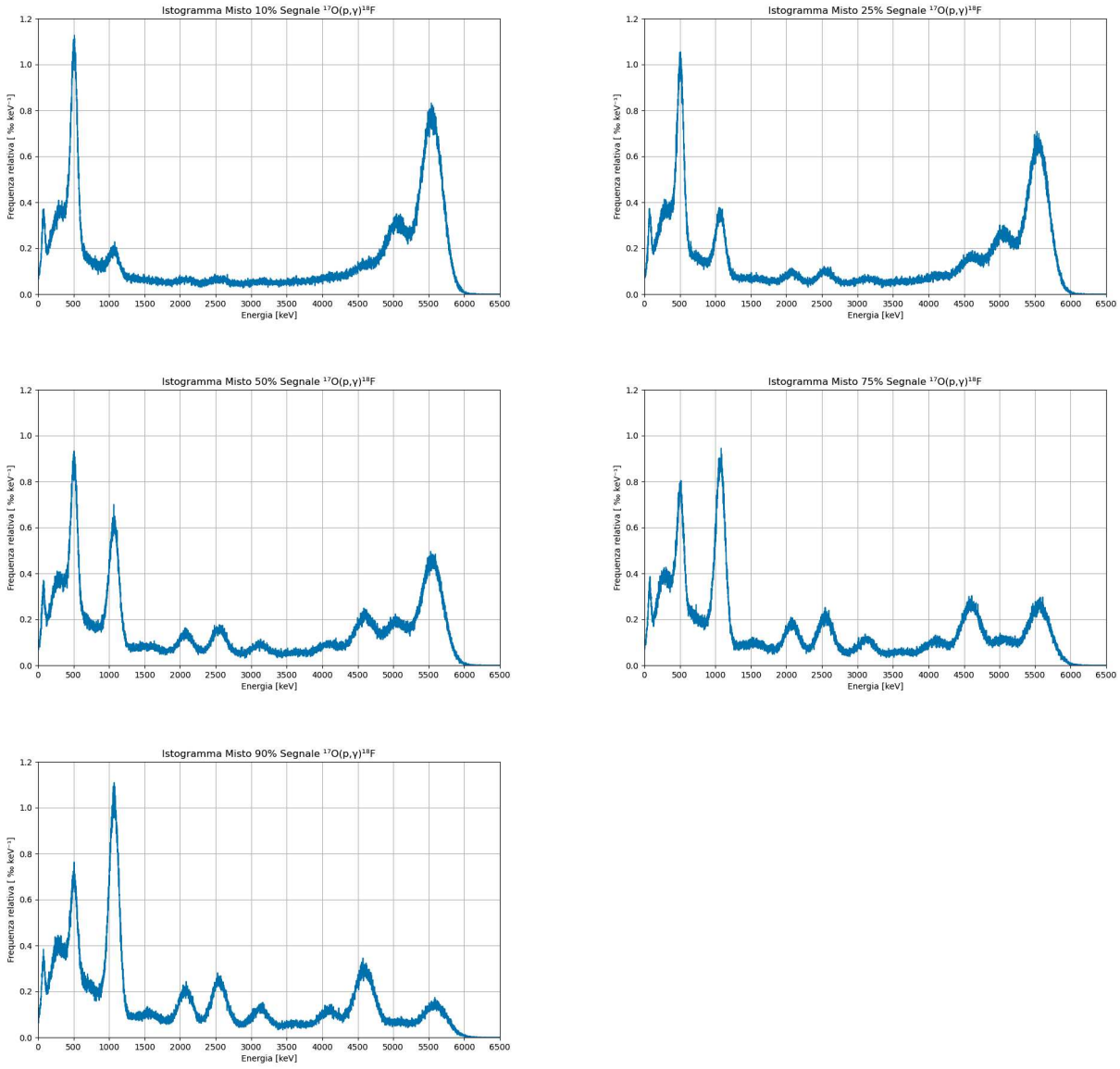

Figura 16: Istogrammi misti segnale-rumore con crescente frazione di $^{17}O(p,\gamma)^{18}F$: all'aumentare di quest'ultima aumentano il numero di picchi a causa della cascata γ della reazione.

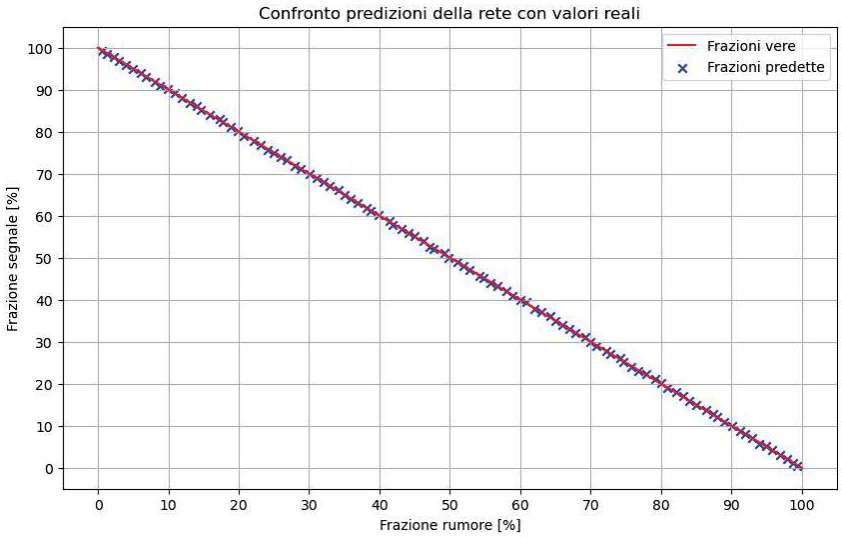

Figura 17: Predizione rete vs dati reali: in rosso la retta con i dati reali con la quale si confrontano i valori determinati dalla rete per proporzioni delle due reazioni.

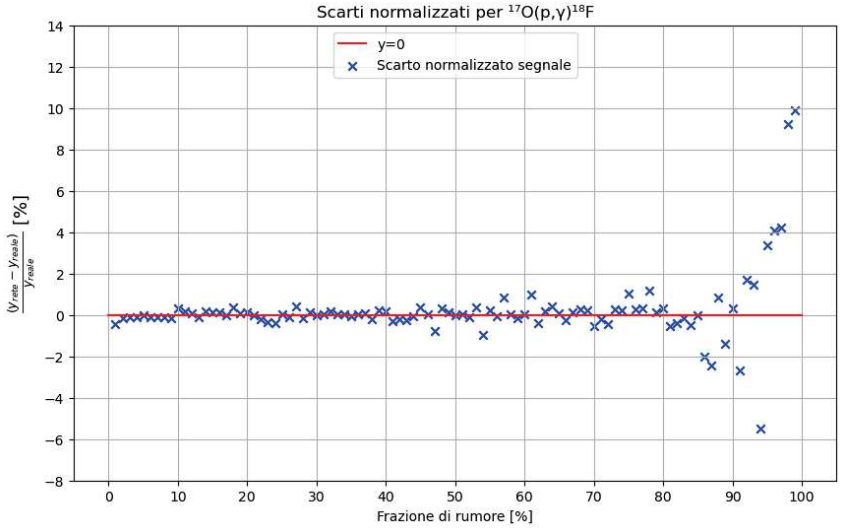

Figura 18: Scarti normalizzati segnale $^{17}O(p,\gamma)^{18}F$.

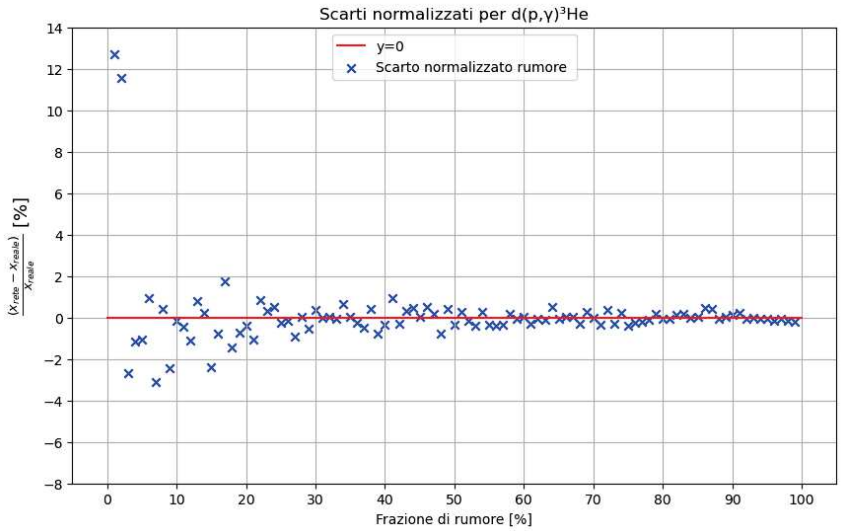

Figura 19: Scarti normalizzati rumore $d(p,\gamma)^3He$.

La rete si comporta molto bene nella regione in cui la frazione di segnale risulta compresa tra 5% e 95%, per cui le predizioni su entrambi segnale e rumore si discostano dai rispettivi valori veri meno del 5%, ottenendo perfino degli errori relativi minori di 1% quando la frazione di segnale è compresa 20% e 80%. Questo fornisce un'ottima predizione in condizioni simili a quelle sperimentali, in cui il contributo di $d(p,\gamma)^3He$ è quello più preponderante. Inoltre, quando il segnale è presente per meno del 5%, la stima sul suo valore risulta più imprecisa, con un errore relativo fino al 10%.

Per quanto riguarda la la regione per cui la porzione di $d(p,\gamma)^3He$ consiste in meno del 5% dello spettro totale, corrispondente sperimentalmente ad una scarsa presenza di contaminante, la predizione della rete sulla frazione di rumore peggiora considerevolmente, con un errore relativo che supera 10%. Questo è dovuto sia al fatto che per, piccoli valori della frazione di rumore, leggere variazioni nella prestazione della rete influiscono in maniera più significativa nella sua predizione (un comportamento analogo lo si ha per il segnale quando la frazione di rumore $\rightarrow 100\%$), sia al fatto che nella rete non si vincoli esplicitamente la somma delle frazioni delle due reazioni all'unità. In queste situazioni è perciò più conveniente basarsi sulla predizione della frazione di $^{17}O(p,\gamma)^{18}F$ e ottenere il contributo di $d(p,\gamma)^3He$ per differenza.

In questa trattazione si sono ignorati contributi al di fuori delle due reazioni considerate, come per esempio il background intrinseco del detector: l'approssimazione è ragionevole, in quanto quest'ultimo arriva solo a 2.5 MeV, e può essere dunque ignorato senza penalizzare la prestazione della rete, allenata supponendo la presenza delle sole $^{17}O(p,\gamma)^{18}F$ e $d(p,\gamma)^3He$.

4 Conclusioni

A LUNA si è studiata la risonanza a 65 keV di $^{17}O(p,\gamma)^{18}F$, riscontrando una contaminazione del segnale dovuta alla reazione con un simile Q-valore, $d(p,\gamma)^3He$, causata dal deuterio presente sulla superficie del target. In questa Tesi si è analizzato l'utilizzo di una rete neurale al fine di distinguere le porzioni delle due reazioni a partire dallo spettro in energia simulato su un cristallo del rivelatore BGO, senza dover, perciò, effettuare un'analisi evento per evento. L'architettura di *Encoder* scelta, una volta ottimizzata aumentando il numero di neuroni negli strati interni, fornisce ottime predizioni sia per la porzione di segnale che di rumore e, quando quest'ultimo risulta inferiore al 5%, conviene privilegiare la stima sul segnale.

La metodologia risulta, dunque, promettente, ed analisi future possono sfruttarla utilizzando, per esempio, dati sperimentali reali. Inoltre, è possibile raffinare il modello introducendo una convoluzione che coinvolga tutti e sei i cristalli disponibili del rivelatore BGO per permettere alla rete di conoscere pattern più complessi.

Riferimenti bibliografici

[1] Gesuè RM (2021), "Direct determination of the $^{17}O(p,\gamma)^{18}F$ 65 keV resonance strenght at LUNA", Laurea Magistrale in Fisica presso Università degli studi di Napoli Federico II.

[2] Mehta P, Wang CH, Day AG, Richardson C, Bukov M, Fisher CK, Schwab DJ (2019), "A high-bias, low-variance introduction to Machine Learning for physicists", *Phys Rep.* 810, pp. 1-124. DOI = https://doi.org/10.1016/j.physrep.2019.03.001

[3] Skowronski J et al (2023), "Advances in radiative capture studies at LUNA with a segmented BGO detector", *J. Phys. G: Nucl. Part. Phys.* 50 045201

[4] Skowronski J (2023), "Neural Networks for Experimental Nuclear Astrophysics: The $^{17}O(p,\gamma)^{18}F$ Resonance at 65 keV as a Test Case", ALPACA Workshop 2023