SAPIENZA
UNIVERSITÀ DI ROMA

Intelligenza Artificiale e Nuova Fisica: Anomaly Detection nei Jet adronici

Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea Triennale in Fisica

Gabriel di Paolo
Matricola 1917882

Relatore
Prof. Stefano Giagu

Anno Accademico 2024-2025

Intelligenza Artificiale e Nuova Fisica: Anomaly Detection nei Jet Adronici
Sapienza Università di Roma

Questa tesi è stata composta con LaTeX e la classe Sapthesis.

Email dell'autore: dipaolo.1917882@studenti.uniroma1.it

Dedicato a chi ha creduto in me più di quanto io abbia mai fatto.

Indice

1 Jet adronici e nuova fisica **2**
1.1 Il Modello Standard delle particelle . . . 2
1.2 Il Large Hadron Collider . . . 4
1.3 Jet adronici da particelle con elevato boost di Lorentz al LHC . . . 6
1.4 Ricostruzione dei jet adronici e studio della loro sottostruttura . . . 7

2 Intelligenza artificiale e anomaly detection **11**
2.1 Introduzione all'Intelligenza Artificiale e al Machine Learning . . . 11
2.1.1 Bias-Variance Tradeoff . . . 13
2.2 Deep Learning e Deep Neural Networks . . . 14
2.2.1 Convolutional Neural Networks . . . 16
2.3 Convolutional Autoencoder e Anomaly Detection . . . 17
2.3.1 Autoencoder e Convolutional Autoencoder . . . 17
2.3.2 CoAE per Anomaly Detection . . . 18

3 Applicazione dell'anomaly detection nell'identificazione dei jet adronici **20**
3.1 Preprocessing del dataset . . . 21
3.2 Definizione del modello CoAE . . . 21
3.2.1 Addestramento del modello . . . 22
3.3 Anomaly detection . . . 22
3.3.1 Metriche utilizzate . . . 23
3.3.2 Caso del quark top . . . 23
3.3.3 Caso del bosone W . . . 24
3.3.4 Caso del bosone Z . . . 24
3.4 Conclusioni . . . 25

Introduzione

In questa dissertazione si espone un approccio che integra il machine learning nella ricerca di segnali di nuova fisica mediante l'identificazione delle anomalie nei dati sperimentali. Nel primo capitolo si introducono il Modello Standard e i jet adronici. Nel secondo capitolo si presentano i concetti base di intelligenza artificiale e machine learning, concentrandosi sulle reti neurali alla base degli autoencoder per la rivelazione delle anomalie. Infine, nel terzo capitolo, viene descritta l'applicazione dell'anomaly detection per identificare specifiche classi di jet adronici.

Capitolo 1

Jet adronici e nuova fisica

1.1 Il Modello Standard delle particelle

Formalizzato attraverso la Teoria dei Campi Quantistici (QFT, Quantum Field Theory) e sviluppato nella seconda metà del ventesimo secolo, il Modello Standard (SM, Standard Model) costituisce la migliore teoria della fisica delle particelle attualmente disponibile e consente ai fisici teorici di fare previsioni molto precise.
Il Modello Standard si basa sul concetto che tutta la materia è fatta di particelle che interagiscono tra loro scambiando altre particelle associate alle forze fondamentali, e tale approccio consente di descrivere come è fatta la materia.
Le particelle del Modello Standard sono suddivise in due gruppi:

- fermioni: particelle elementari con spin semi-intero ($s = \frac{1}{2}\hbar$)[1], che ubbidiscono alla statistica di Fermi-Dirac;
- bosoni: particelle elementari con spin intero ($s = \hbar$), che ubbidiscono alla statistica di Bose-Einstein;

I fermioni si suddividono a loro volta in due gruppi composti da sei quark, che sono soggetti all'interazione forte, e sei leptoni, che invece non risentono dell'interazione forte; ciascun gruppo comprende tre generazioni:

- quark: up (u) e down (d), charm (c) e strange (s), top (t) e bottom (b);

$$\begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix}$$

- leptoni: neutrino elettronico (ν_e) ed elettrone (e), neutrino muonico (ν_μ) e muone (μ), neutrino tauonico (ν_τ) e tauone (τ).

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix} \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$$

Ogni generazione comprende coppie di particelle con massa maggiore rispetto alla generazione precedente, e ai membri di ogni famiglia si può assegnare un particolare

[1] $\hbar = \frac{h}{2\pi}$ è la costante di Planck ridotta, con $h = 6.62607015 \cdot 10^{-34}\, J \cdot s$.

numero quantico, diverso per ogni generazione, che regola le interazioni.
Le particelle composte da quark si definiscono adroni e si distinguono le particelle formate da un numero dispari di quark (barioni, come il protone e il neutrone, hanno spin semi-intero), e quelle formate da un numero pari di quark (mesoni, come il pione e il kaone, hanno spin intero).
I bosoni, invece, sono associati alle forze fondamentali attraverso le quali interagiscono le particelle, e si suddividono in:

- gluoni (g): mediatori della forza nucleare forte che lega i quark negli adroni; sono sprovvisti di massa e di carica elettrica.
- fotoni (γ): mediatori della forza elettromagnetica; sono sprovvisti di massa e di carica elettrica.
- bosoni Z^0 e $W^{\pm}$: mediatori della forza nucleare debole, responsabile della radioattività. Il bosone Z ha carica elettrica nulla, mentre i bosoni W hanno carica elettrica pari a $\pm e$ ($|e| = 1.602176634 \cdot 10^{-19}\, C$, carica elementare), quindi partecipano in modo differente ai processi di decadimento.
- bosone di Higgs (H): mediatore del campo di Higgs, conferisce massa alle particelle; è sprovvisto di carica elettrica.

Il Modello Standard è stato testato da molti esperimenti e si è dimostrato particolarmente efficace nell'anticipare l'esistenza di particelle precedentemente sconosciute. Tuttavia, non include la gravità e non riesce a spiegare perché quest'ultima è molto più debole delle altre interazioni fondamentali. Non fornisce una previsione per le masse dei neutrini e non spiega nemmeno la materia oscura e perché è circa cinque volte più abbondante nell'universo rispetto alla materia ordinaria.
Tutto ciò lascia presupporre che il Modello Standard è ancora incompleto. Pertanto, le informazioni da nuovi esperimenti al Large Hadron Collider (LHC) sono cruciali per scoprire effetti di Nuova Fisica oltre il Modello Standard (BSM, Beyond Standard Model) che potrebbero completare il quadro della teoria.

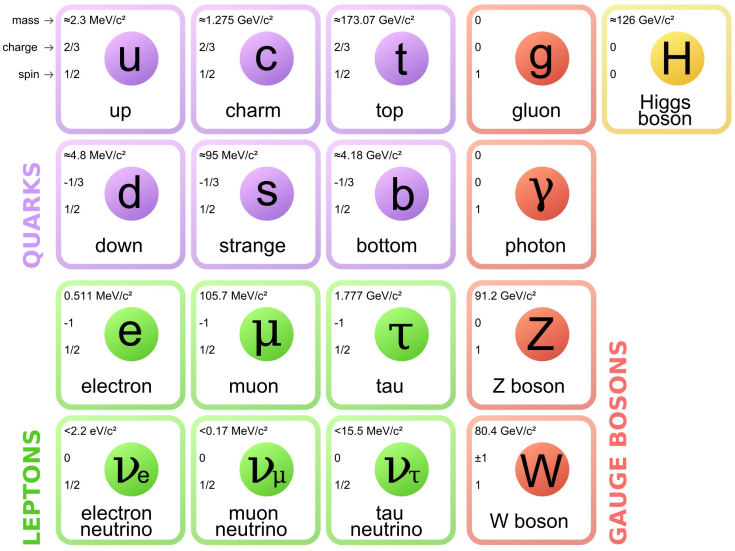

Figura 1.1. Modello Standard delle particelle

1.2 Il Large Hadron Collider

Il complesso di acceleratori del CERN (Consiglio Europeo della Ricerca Nucleare) è formato da una successione di apparati, ciascuno dei quali trasferisce il fascio di particelle da accelerare nel successivo, aumentandone l'energia; il Large Hadron Collider è l'ultimo elemento di questa successione.
Si tratta di un anello di circa 27 km di circonferenza che riesce ad accelerare fasci di ioni o protoni viaggianti in direzioni opposte all'energia record di 6.5 TeV, raggiungendo dunque nella collisione un'energia del centro di massa pari a 13 TeV. Questo è possibile grazie alle sue dimensioni e al sistema di dipoli magnetici ottenuti da elettromagneti conduttori che, lavorando a una temperatura di 1.9 K con una corrente di 11850 A, riescono a generare un campo magnetico di 8.3 T che accelera le particelle lungo la traiettoria circolare.
Facendo un calcolo approssimativo, ogni fascio è composto da circa 3000 pacchetti di particelle che possono contenere ciascuno fino a 100 miliardi di particelle; siccome la probabilità che due particelle si scontrino è molto bassa, quando i pacchetti si incrociano avvengono fino a 40 collisioni tra 200 miliardi di particelle. I pacchetti si incrociano in media circa 40 milioni di volte al secondo, quindi, l'LHC genera circa 1 miliardo di collisioni tra particelle al secondo.

Per catturare quanta più fisica possibile, dal bosone di Higgs agli effetti di nuova fisica, si utilizzano al LHC quattro rivelatori di particelle posti nei quattro punti di collisione dei fasci, dei quali il rivelatore ATLAS (A Toroidal LHC ApparatuS) è il più grande mai costruito. É disposto in modo da avvolgere concentricamente il punto di collisione e registrare la traiettoria, la quantità di moto e l'energia delle particelle, applicando un campo magnetico per deviarle. Questo è possibile grazie a più strati di rivelatori che si compenetrano e che sono costituiti da strati di materiale assorbente che fermano le particelle in arrivo, intervallati da strati di un mezzo attivo che ne misurano diverse quantità cinematiche, tra le quali quantità di moto, energia, direzione e carica. Come si può vedere nello schema seguente, ciascun rivelatore esegue un compito differente.

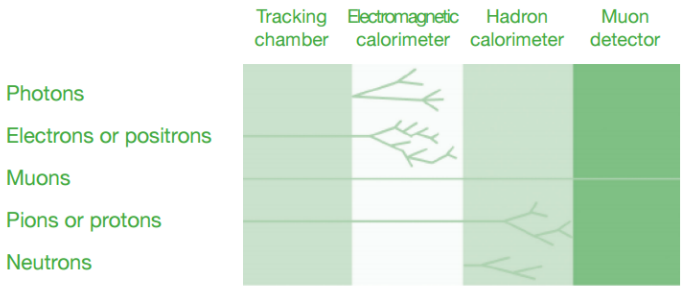

Figura 1.2. Sistema di rivelatori di ATLAS

- Sistema di tracciamento interno: immerso in un campo magnetico parallelo all'asse del fascio, è la prima parte di ATLAS a vedere i prodotti di decadimento delle collisioni; misura la direzione, la quantità di moto e la carica delle particelle prodotte.

- Calorimetro elettromagnetico ad argon liquido: misura l'energia di elettroni e fotoni. È dotato di strati di metallo che assorbono le particelle in arrivo, convertendole in una cascata di nuove particelle a energia inferiore che, ionizzando l'argon, producono corrente elettrica misurabile dalla quale stimare l'energia della particella incidente.
- Calorimetro adronico: misura l'energia degli adroni mediante strati di acciaio e piastrelle scintillanti in plastica. Quando gli adroni colpiscono gli strati di acciaio, generano una pioggia di nuove particelle che, colpendo gli scintillatori, producono fotoni che vengono convertiti in corrente elettrica con intensità proporzionale all'energia della particella originale.
- Spettrometro a muoni: identifica e misura il momento dei muoni che attraversano gli strati precedenti del rivelatore.

Per facilitare l'analisi di eventi e la visualizzazione delle interazioni particellari, si adotta un sistema di coordinate polari, con il quale si può mappare ogni particella in un piano che rappresenta tutte le direzioni attorno al punto di collisione, sfruttando la simmetria del rivelatore. Questo sistema di coordinate ha l'origine centrata nel punto di collisione dell'esperimento, l'asse y che punta verticalmente verso l'alto, l'asse x che punta radialmente verso il centro del LHC e l'asse z che punta verso la direzione dei fasci. Si definisce il piano trasverso xy e r la coordinata radiale in questo piano. Identificando un angolo azimutale ϕ, misurato rispetto all'asse x nel piano trasverso, e un angolo polare θ, misurato rispetto all'asse z, si definisce la pseudorapidità come $\eta = -ln(tan(\theta/2))$, che rappresenta l'angolo di una particella rispetto all'asse del fascio; per $\eta \to \infty$ la particella si muove quasi parallela all'asse del fascio ($\theta \to 0°, 180°$), mentre per $\eta \to 0$ la particella si muove con un angolo più ampio rispetto all'asse z ($\theta \to 90°$). In questo sistema di coordinate, la distanza tra punti si può così definire: $\Delta R^2 = \Delta\eta^2 + \Delta\phi^2$.
Per lo studio della cinematica delle particelle, è conveniente definire nel piano trasverso la quantità di moto trasversa $p_T = |p|sen\theta$ e l'energia trasversa $E_T = |E|sen\theta$, sulla base delle quali si definisce il momento trasverso mancante come somma vettoriale di tutti i momenti trasversi: $E_T^{miss} = |\sum_i p_{T,i}|$. Per la conservazione del momento, questa quantità dovrebbe essere nulla; tuttavia, si potrebbe misurare uno sbilanciamento quando una nuova particella sfugge ai rivelatori.

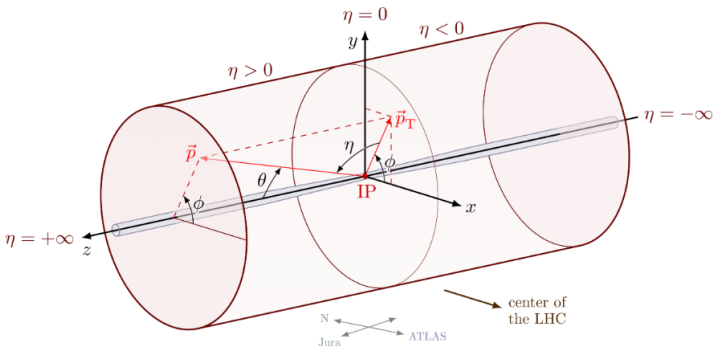

Figura 1.3. Sistema di coordinate polari

1.3 Jet adronici da particelle con elevato boost di Lorentz al LHC

Durante gli esperimenti di fisica delle particelle al LHC, in particolare durante le collisioni protone-protone, gli oggetti rilevati più di frequente sono fasci altamente collimati di adroni che prendono il nome di jet adronici.

Quando i protoni collidono ad alte energie, si originano molti partoni energetici, oggi comunemente chiamati quark e gluoni; infatti, i protoni sono formati da tre quark (due quark up e un quark down) e un numero indeterminato di gluoni. Queste particelle non possono essere rivelate direttamente, ma danno origine a jet di particelle osservabili. Infatti, i partoni energetici decadono a loro volta in gluoni e coppie quark-antiquark, generando un fascio collimato di partoni detto parton-shower, e quando raggiungono il raggio di confinamento del colore (circa 1 Å, distanza oltre la quale il colore non può esistere liberamente) adronizzano, ovvero si combinano a formare gli adroni stabili che costituiscono il jet adronico. Questi jet costituiscono depositi di energia macroscopicamente rilevabili dei quark e dei gluoni prodotti dallo scattering protone-protone.

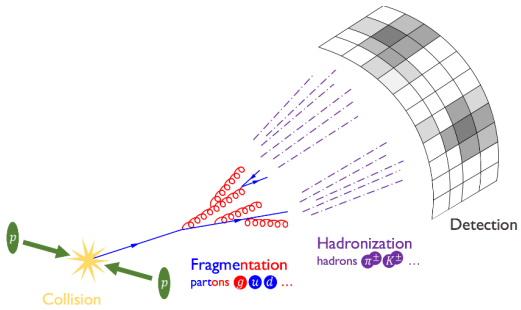

Figura 1.4. Creazione del jet adronico dallo scattering protone-protone e conseguente deposito di energia nei calorimetri

Il Large Hadron Collider, come detto, studia la fisica sulla scala del Teraelettronvolt (TeV); grazie all'energia in gioco, durante le collisioni più energetiche, le particelle più pesanti note, il bosone W ($m_W = 80.4\,GeV/c^2$), il bosone Z ($m_Z = 91.2\,GeV/c^2$), il bosone di Higgs ($m_H \approx 126\,GeV/c^2$) e il quark top ($m_t \approx 173.07\,GeV/c^2$), vengono prodotte con elevata energia cinetica. Queste particelle pesanti non vengono osservate di per sé, ma decadono in altre particelle che vengono catturate dai rivelatori.
Questo tipo di collisioni energetiche permette uno studio particolare dei prodotti di decadimento, poiché a seconda della velocità della particella che decade, cambia il modo in cui si rivelano i suoi prodotti di decadimento. Infatti, se la particella che decade è a riposo, allora il decadimento sarà back-to-back; mentre se la particella che decade è altamente energetica, allora la topologia del decadimento è diversa e si rivelano i prodotti del decadimento emessi nella stessa direzione, al punto che per un grande boost di Lorentz i prodotti del decadimento vengono emessi a piccoli angoli rispetto alla direzione originale della particella che decade.
Dunque, quando particelle con grande boost di Lorentz decadono in coppie quark-

antiquark, i jet adronici che ne derivano saranno sovrapposti a formare un unico jet detto fat-jet.

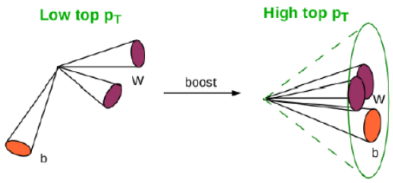

Figura 1.5. Semplice schema del decadimento del quark top in una configurazione poco energetica e in una configurazione con grande boost di Lorentz.

Ai fini della trattazione è bene distinguere tra diversi tipi di jet adronici. Infatti, oltre ai jet adronici prodotti direttamente da quark e gluoni generati nelle collisioni protone-protone, che rappresentano fenomeni ordinari governati dalle interazioni forti, è possibile rivelare anche jet prodotti dal decadimento di particelle instabili quali i bosoni W, Z, e quark top.
I bosoni W e Z hanno vari canali di decadimento, tra i quali possono decadere in quark che a loro volta generano jet adronici. Il bosone W può decadere in coppie di quark $u + \bar{d}$, mentre il bosone Z può decadere in una coppia quark-antiquark di qualsiasi tipo; quindi si rivelano sempre due jet molto vicini originati dai due quark quando i bosoni che decadono sono molto energetici.
Il quark top, invece, decade quasi istantaneamente in $W + b$, con W che può a sua volta decadere in altri quark; di conseguenza si possono osservare ben tre jet, uno prodotto dal quark bottom e due dal bosone W, che potrebbero essere fusi.

Lo studio dei jet adronici è di grande importanza per la scoperta di nuove particelle. Infatti, le caratteristiche di un jet, come la sua energia o la distribuzione del momento angolare dei suoi costituenti, riflettono la configurazione altamente energetica iniziale delle particelle che lo hanno generato.
Inoltre, negli scenari BSM ci sono risonanze pesanti che decadono in particelle con massa intermedia, che a loro volta decadendo formano jet, secondo il processo $X \to Y \to jet$, con X risonanza sconosciuta e Y che può essere una tra le particelle W, Z, t oppure una particella oltre il Modello Standard; in genere, quando $m_X >> m_Y$ la particella intermedia ha un grande boost di Lorentz.
Dunque, distinguere questi tipi di jet può portare alla scoperta di nuove particelle oltre il Modello Standard, poiché molte teorie di Nuova Fisica prevedono la produzione di nuove particelle che decadono in modo simile alle particelle W, Z, t ma con differenze sottili.

1.4 Ricostruzione dei jet adronici e studio della loro sottostruttura

Quando si analizzano i dati provenienti dalle collisioni protone-protone, si prendono in considerazione miliardi di eventi per volta; dunque, affinché l'analisi sia ottimale

è necessario dare una buona definizione computazionale di jet adronico, che possa essere sfruttata dagli algoritmi di ricostruzione dei jet.

La ricostruzione del jet adronico avviene raccogliendo prima di tutto i dati provenienti dai depositi di energia nei calorimetri che circondano il punto di collisione protone-protone. Questi calorimetri, come detto, misurano l'energia delle particelle incidenti assorbendo la cascata di particelle prodotta nel materiale al passaggio della particella, e sono posti parallelamente all'asse dei fasci incidenti, così da poter catturare il passaggio di particelle che si propagano radialmente dal punto di collisione. La luce generata dalle scintillazioni viene trasferita nella regione del visibile mediante delle fibre che variano la lunghezza d'onda, raccolgono la luce e la convogliano verso i rivelatori.

Con i dati provenienti dai calorimetri è possibile ricostruire in tre dimensioni le regioni di energia depositata dalle particelle all'interno del volume dei calorimetri; mediante algoritmi di clustering topologico, è possibile suddividere tali regioni in topo-cluster, ovvero raggruppamenti di celle calorimetriche con energia significativa; l'elevata granularità dei rivelatori calorimetrici consente ai topo-cluster raccolti di includere informazioni dettagliate relative al flusso di energia dei jet.
A questo punto si possono ricostruire i jet mediante gli algoritmi di ricostruzione, che seguono il principio di funzionamento seguente. Prima di tutto, per uno specifico cluster i si definiscono le quantità $d_{iB} = p_{T,i}^{2k}$ (distanza del cluster i dal fascio B) e $d_{ij} = min(d_{iB}, d_{jB}) \cdot \frac{\Delta R_{ij}^2}{R^2}$ (distanza del cluster i dal cluster j), con p_T momento trasverso associato al fascio del cluster, e $\Delta R_{ij}^2 = \Delta\eta^2 + \Delta\phi^2$. R è un parametro che regola il raggio del jet e la sua ricostruzione, ovvero determina la distanza limite entro la quale due particelle possono essere considerate collineari e quindi appartenenti a un unico jet; infatti, riducendo questo parametro, l'identificazione dei jet diventa più fine, dunque, il numero di jet in un evento dipende direttamente da questo parametro. k, invece, è il termine che definisce il tipo di algoritmo utilizzato per la ricostruzione: $k = 1$ per l'algoritmo k_t, $k = -1$ per anti-k_t e $k = 0$ per l'algoritmo Cambridge-Aachen. Definite queste quantità, si fa un'analisi di tutte le distanze d_{iB} e d_{ij} cercando la distanza minima d_{min}; se d_{min} fa parte delle d_{iB}, allora il cluster è dichiarato jet e rimosso dalla lista, mentre se d_{min} fa parte delle d_{ij} i cluster i e j sono fusi. Questo procedimento viene ripetuto fino ad esaurire la lista dei cluster.

Tra l'algoritmo k_t e anti-k_t ci sono alcune differenze importanti che riguardano la fusione dei cluster. Infatti, l'algoritmo k_t unisce per primi i cluster a bassa energia, quindi risulta molto sensibile alla presenza di particelle a bassa energia adiacenti; anche se questo garantisce una certa somiglianza tra il jet teorico e quello ricostruito, quando gli effetti di contaminazione sono molto maggiori (come avviene al LHC), i jet possono assumere forme asimmetriche. Per questo motivo viene usato più comunemente l'algoritmo anti-k_t, che raggruppa per prime le componenti più energetiche, conferendo ai jet una forma circolare ben definita.

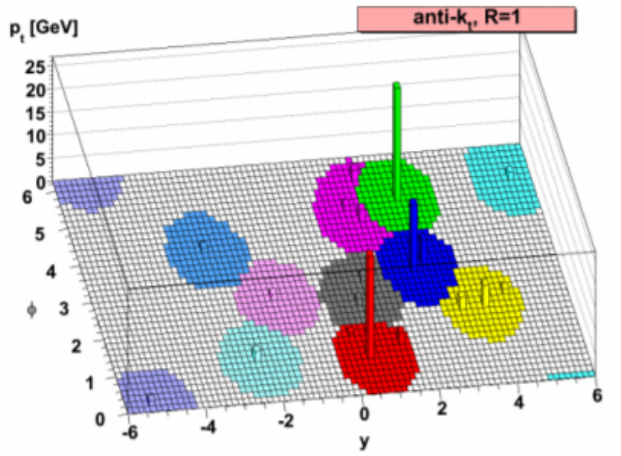

Figura 1.6. Esempio di jet adronici ricostruiti con l'algoritmo anti-k_t.

Alcuni tra i problemi che rendono instabili gli algoritmi di ricostruzione dei jet sono la contaminazione dovuta al pileup e agli eventi di fondo.
Quando avviene lo scattering protone-protone, infatti, si considera soltanto l'urto tra un partone nel primo protone e un partone nel secondo protone; tuttavia, è possibile che più di un solo partone per protone prenda parte allo scattering, e di solito questa interazione produce particelle a bassa energia (poiché la probabilità di tali eventi è molto bassa). Inoltre, quando si eseguono esperimenti di fisica delle particelle negli acceleratori come LHC, i fasci di protoni che collidono contengono un gran numero di protoni, pertanto possono esserci più di un singolo urto che avvengono simultaneamente, anche se la probabilità di questi eventi secondari è piuttosto bassa. Aumentando l'energia del centro di massa, aumenta la probabilità degli urti secondari e di eventi di fondo, che quindi vanno a complicare la ricostruzione del jet.
L'algoritmo, inoltre, deve risultare IRC safe (InfraRed and Collinear safe), cioè deve essere stabile rispetto a emissioni di gluoni a bassa energia e deve essere tale che quando due particelle sono collineari non ci siano cambiamenti nel jet.
Inoltre, sorge un problema legato al parametro R nel caso dei prodotti di decadimento con grande boost di Lorentz. In genere, per il decadimento di particelle massive accelerate, si può fare la seguente stima per il raggio della regione di decadimento: $\Delta R = \frac{2m}{p_T}$, utilizzando i dati relativi alla particella che decade. Se questo raggio è più piccolo del raggio impostato per la ricostruzione del jet, allora l'algoritmo ricostruisce i prodotti collimati come un singolo jet di raggio R che viene definito fat-jet. Quindi, quando si analizzano sistemi con alto boost di Lorentz, gli algoritmi standard di ricostruzione dei jet falliscono a causa della forte collimazione dei prodotti di decadimento in una regione con raggio minore di R.

Tuttavia, c'è un modo per differenziare i jet prodotti da particelle pesanti boosted che decadono da quelli prodotti invece da gluoni e quark. Mentre i fat-jet hanno una sottostruttura caratteristica generata dai jet collimati, lo stesso non si può dire per i jet generati da quark e gluoni, nei quali invece si può vedere una zona centrale più energetica circondata da una regione a bassa energia dovuta al pileup e agli eventi di fondo. In altre parole, nel primo caso si possono trovare cluster all'interno

del jet ricostruito nelle zone a più alta energia, mentre nel secondo caso questo non è possibile. Per facilitare la distinzione tra questi jet si utilizzano tecniche che rimuovono la radiazione associata al pileup e agli eventi di fondo; in ATLAS si usano i seguenti algoritmi: mass-drop filtering, trimming e pruning.

Il mass-drop filtering opera sui singoli cluster che formano il jet ricostruito e isola quelli che hanno una massa molto minore rispetto al jet originario, così da scartare cluster che rappresentano fluttuazioni casuali piuttosto che reali strutture di decadimento.
Il jet trimming, invece, opera sui singoli cluster che formano il jet ricostruito tramite k_t e compara i loro momenti trasversi con quello del jet ricostruito; si basa sul fatto che la contaminazione dovuta al pileup e agli eventi di fondo è di bassa energia, pertanto scarta quei cluster i cui momenti trasversi non soddisfano la condizione $p_T^i > f_{cut} \cdot \Lambda_{hard}$, con f_{cut} parametro dell'algoritmo e Λ_{hard} parametro che governa la cinematica dei processi forti.
Il jet pruning, infine, segue lo stesso principio del jet trimming ma opera in fase di ricostruzione del jet; infatti, quando si trova una d_{min}, invece di unire subito i due cluster, si richiede che il loro momento trasverso rispetti la condizione $\frac{min(p_T^i, p_T^j)}{p_T^{i+j}} > z_{cut}$ e che $\Delta R_{ij} < D_{cut}$, con z_{cut} e D_{cut} parametri dell'algoritmo. Numerose analisi propongono come valori ottimali $z_{cut} = 0.1$ e $D_{cut} = \frac{m^i}{p_T^i}$.

Inoltre, le tecniche moderne di analisi dati si stanno focalizzando sullo studio della sottostruttura dei jet adronici. In particolare, per categorizzare i jet in diverse tipologie, è possibile definire un insieme di variabili che descrivono la loro sottostruttura. Tali variabili sono strettamente legate sia alla cinematica del jet, sia alle sue caratteristiche topologiche. Come detto, lo scopo principale è quello di differenziare i jet adronici prodotti da quark e gluoni rispetto a quelli prodotti da W, Z e top.
I jet prodotti da quark e gluoni ad alta energia hanno una sottostruttura complessa che è importante per studiare il Modello Standard e che fornisce lo strumento per sondare la natura alle scale di energia più elevate; inoltre, il tagging dei jet di quark e gluoni è interessante perché molti stati finali di interesse SM e BSM sono dominati da processi di fondo con jet e gluoni. I bosoni W e Z spesso partecipano a scenari BSM, quindi possono essere presenti in molti stati finali di questi modelli. La stessa cosa si può dire del quark top; ci sono molti scenari BSM che hanno quark top altamente energetici che partecipano agli eventi.

Capitolo 2

Intelligenza artificiale e anomaly detection

2.1 Introduzione all'Intelligenza Artificiale e al Machine Learning

Al giorno d'oggi l'intelligenza artificiale (IA) è un campo in piena espansione che ha molte applicazioni pratiche; infatti si utilizzano quotidianamente software intelligenti per facilitare i lavori di routine, comprendere immagini o tradurre linguaggi, fare diagnosi mediche e supportare la ricerca scientifica in generale.
I modelli di intelligenza artificiale prendono in input dei dati, descritti da variabili (dette feature) che ne descrivono le proprietà distintive, con il compito di analizzarli e acquisire conoscenza estraendola direttamente da essi; questo approccio all'IA viene detto machine learning (ML), e conferisce ai modelli la capacità di imparare dai dati e generalizzare.

E' possibile differenziare i modelli di machine learning in base all'approccio con il quale apprendono dai dati. Infatti, l'apprendimento si dice supervisionato quando ad ogni esempio $\mathbf{x}$ è associato un set di valori $\mathbf{y}$ e il modello impara a stimare la probabilità condizionata $p(\mathbf{y}|\mathbf{x})$ per predire $\mathbf{y}$ da $\mathbf{x}$; quindi, per ogni esempio $\mathbf{x}$ è definito un target $\mathbf{y}$. Invece, nell'apprendimento non supervisionato, non viene specificato un target per gli esempi in input; quindi, il modello cerca di stimare l'intera distribuzione di probabilità del dataset $p(\mathbf{x})$ apprendendo dai dati stessi senza ulteriori indicazioni.

Il processo di apprendimento del modello può essere riassunto dalla definizione data da Tom Mitchell nel 1997: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."
Quando si parla di task (T), si intende l'obiettivo specifico che il modello deve raggiungere nel processare gli esempi, ovvero un insieme di feature che sono state misurate per l'evento che vogliamo studiare; generalmente, ciascuna feature x_i rappresenta una proprietà osservabile dell'esempio ed è un elemento di un vettore $\mathbf{x} \in \mathbb{R}^n$, con n numero totale delle feature. Tra i task più comuni ci sono la

classificazione, la regressione e l'anomaly detection. La classificazione associa ogni input $\mathbf{x}$ a una specifica classe, identificata da un'etichetta nell'insieme $\{1, ..., k\}$, attraverso la definizione di una funzione $f : \mathbb{R}^n \to \{1, ..., k\}$, tale per cui se $y = f(\mathbf{x})$, allora l'input appartiene alla classe y. La regressione, invece, predice un valore numerico in output dato un valore in input, tramite una funzione $f : \mathbb{R}^n \to \mathbb{R}$. L'anomaly detection, infine, richiede al modello di analizzare una serie di dati in input per identificare elementi che si discostano significativamente dal comportamento atteso o normale; questo compito si basa generalmente su un modello che rappresenta il comportamento normale dei dati e assegna un punteggio di anomalia a ciascun esempio, classificando come anomalie i dati con punteggi elevati.

L'abilità del modello di machine learning di rispondere al task proposto viene determinata attraverso una misura delle sue performance (P); data la grande varietà di task, non è sempre facile scegliere secondo quale criterio testare le performance del modello. Nel caso della classificazione, per esempio, una misura della capacità di discriminare le classi da parte del modello può avvenire attraverso l'accuracy, ovvero la quantità di esempi che sono stati classificati bene, oppure in maniera analoga, attraverso l'error rate, ovvero la quantità di esempi che sono stati classificati erroneamente.
Il criterio secondo il quale si monitora l'apprendimento del modello è la funzione di perdita o loss function, che calcola la distanza tra il valore predetto dal modello e il valore vero dell'output; l'obiettivo è quindi ottimizzare questa funzione. Tra le funzioni di perdita utilizzate più spesso ci sono:

- Mean Absolute Error $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$: calcola la media dell'errore assoluto tra il valore vero e il valore predetto, fornendo una misura molto sensibile agli outlier.
- Mean Squared Error $MSE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|^2$: calcola la media del quadrato dello stesso errore, fornendo una misura che penalizza maggiormente gli errori grandi; a differenza della MAE, la MSE è differenziabile.
- Binary Cross Entropy $BCE = -\frac{1}{n}\sum_i [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$: spesso utilizzata per modelli con output in $(0, 1)$, valuta la differenza tra la distribuzione di probabilità predetta e quella reale, introducendo una penalizzazione logaritmica per predizioni molto lontane dai valori veri.

Generalmente sono funzioni $L : \mathbb{R}^n \to \mathbb{R}$, e una tecnica attraverso la quale possono essere minimizzate è quella della discesa lungo il gradiente (Cauchy, 1847), che cerca il minimo della funzione variando i parametri con piccoli step nel verso opposto al segno della derivata: $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla_{\mathbf{w}} L(\mathbf{w}_k)$, dove η è uno scalare positivo detto learning rate che determina la dimensione dello step e k è l'indice dell'iterazione. La minimizzazione converge quando ogni elemento del gradiente è prossimo a zero. Tuttavia, il costo computazionale di questa operazione cresce linearmente con la dimensione del dataset analizzato, e non sempre converge verso minimi globali. Pertanto, quando si lavora con dataset di grandi dimensioni, si ricorre all'algoritmo di discesa stocastica lungo il gradiente (SGD, Stochastic gradient descent), nel quale il gradiente rappresenta un valore atteso, e tale aspettativa può essere approssimata utilizzando un piccolo sottoinsieme di campioni, ovvero un minibatch di esempi scelti casualmente in modo uniforme dal training set. Il campionamento casuale

degli eventi nel minibatch introduce una fonte di rumore, presente anche una volta raggiunto il minimo; questo rende necessario diminuire gradualmente il learning rate per garantire la convergenza. La scelta del valore del learning rate è cruciale per l'apprendimento. Infatti, se è troppo grande, l'apprendimento non è stabile, e la funzione di costo può aumentare significativamente. Invece, se il learning rate è troppo basso, l'apprendimento potrebbe bloccarsi con un valore di costo elevato oppure potrebbe procedere lentamente.

Alcuni algoritmi sono in grado di adattare il learning rate durante l'addestramento del modello. L'algoritmo AdaGrad (Adaptive Gradient) regola il learning rate in base alla somma dei gradienti al quadrato accumulati fino a quel momento; è molto efficace per funzioni convesse, ma per funzioni non convesse può ridurre troppo velocemente il learning rate, rallentando l'addestramento. L'algoritmo RMSProp (Root Mean Square Propagation) modifica l'algoritmo AdaGrad; invece di considerare tutti i gradienti passati, utilizza una media che decresce esponenzialmente per accumulare i gradienti passati, e questo gli consente di essere più veloce e più efficiente anche per funzioni non convesse.

Nel 1964 Polyak introduce il metodo del momento per accelerare l'ottimizzazione dei modelli. Risulta molto utile quando ci sono zone con curvature elevate, oppure quando i gradienti sono piccoli o rumorosi; quindi, quando il gradiente cambia rapidamente in una direzione piuttosto che nelle altre, quando porta lentamente nella direzione giusta o quando oscilla rapidamente. L'idea consiste nel salvare una media esponenzialmente decrescente dei gradienti precedenti per continuare a muoversi nella loro direzione; in questo modo, l'aggiornamento avviene mediante un nuovo parametro $\mathbf{v}$ che funge da velocità: $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \mathbf{v}_{k+1}$, con $\mathbf{v}_{k+1} = \alpha \mathbf{v}_k + (1-\alpha)\nabla L(\mathbf{w}_k)$. Se $\alpha = 0$ si ottiene la discesa lungo il gradiente convenzionale, mentre se $\alpha = 1$ si ignora la discesa lungo il gradiente e si procede nella direzione dello step precedente.

Unendo RMSProp e il metodo del momento nasce l'algoritmo Adam (adaptive momentum) che è l'ottimizzatore più utilizzato nell'addestramento dei modelli di machine learning.

2.1.1 Bias-Variance Tradeoff

La misura delle prestazioni del modello viene eseguita su un set di dati a parte detto test set; infatti, lo scopo è quello di testare la generalizzazione del modello, ovvero la sua capacità di ottenere buoni risultati su dati mai visti in fase di addestramento. In questa fase possono verificarsi due situazioni: l'errore sul test set è molto maggiore dell'errore sul training set (overfitting), oppure l'errore sul training set non viene ottimizzato (underfitting). Tali comportamenti sono modulati dalla complessità del modello e dalla sua capacità, ovvero la sua abilità di apprendere una grande varietà di funzioni. Infatti, un modello con capacità limitata potrebbe non riuscire ad apprendere correttamente in fase di addestramento, quindi non è in grado di ottimizzare l'errore sul training set; invece, un modello con capacità più ampia potrebbe adattarsi troppo agli esempi visti in fase di addestramento, al punto di

non essere in grado di generalizzare sui dati di test.

Per comprendere l'errore di generalizzazione del modello è bene definire i concetti di bias e varianza. Il bias è una misura della differenza tra il valore vero e il valore medio predetto dal modello, mentre la varianza è una misura di quanto le stime del modello cambiano se cambiano i dati di allenamento. Modelli troppo semplici hanno bias alto e varianza bassa, tendendo a generalizzare male perché non catturano abbastanza la complessità del problema. Invece, modelli troppo complessi hanno bias basso e varianza alta, con conseguente adattamento eccessivo ai dati di addestramento e quindi poca capacità di generalizzazione su nuovi dati. Il compromesso (Bias-Variance Tradeoff) è dato da un modello che regola bias e varianza in modo da ottenere una buona capacità di generalizzazione.
In genere, oltre al training set e al test set, si definisce anche un validation set, sul quale ottimizzare gli iperparametri del modello sulla base dell'errore di generalizzazione.

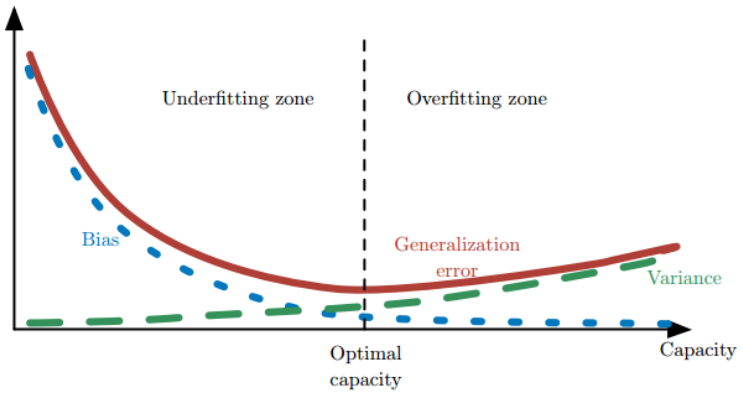

Figura 2.1. Aumentando la capacità del modello, diminuisce il bias ma aumenta la varianza; la capacità ottimale è tale da minimizzare l'errore di generalizzazione.

2.2 Deep Learning e Deep Neural Networks

Gran parte della riuscita del processo di apprendimento dipende dalla rappresentazione dei dati forniti, in particolare dalle features che ne descrivono l'informazione. Non sempre si conoscono a priori le features più importanti da estrarre dai dati per avere delle buone performance su uno specifico compito; pertanto, è necessario un approccio al machine learning detto representation learning, che permette al modello di trovare un buon insieme di features di alto livello che spieghino i dati stessi. Questo è possibile grazie al deep learning (DL), che permette al modello di apprendere concetti e disporli secondo una gerarchia, così da costruire concetti complicati partendo da concetti più semplici. Il miglior esempio, in questo caso, è dato dal feedforward deep network, detto anche multilayer perceptron (MLP), ovvero una rete profonda che approssima una funzione matematica in grado di mappare l'input nell'output, partendo da funzioni più semplici. La rete neurale feedforward prende il nome dal fatto che l'informazione fluisce in avanti, dall'input all'output, senza feedback. E' definita rete perché è composta da più funzioni semplici, organizzate in

strati il cui numero ne definisce la profondità. Infine, è detta neurale perché ispirata al funzionamento dei neuroni biologici.

L'organizzazione a strati della rete neurale artificiale prevede un primo strato detto input layer e uno strato finale detto output layer, tra i quali ci sono gli hidden layers, chiamati così perché il loro output non è visibile; ogni strato è funzione di quello precedente e sono presenti tutte le connessioni possibili tra strati (fully connected layers).
Di solito questi modelli devono apprendere funzioni non lineari; a tal proposito le feedforward networks soddisfano il Teorema dell'Approssimazione Universale (Hornik et al., Cybenko, 1989), che afferma che una rete feedforward con un output lineare e almeno un hidden layer può approssimare qualsiasi funzione continua in un sottoinsieme chiuso e limitato di $\mathbb{R}^n$.

I neuroni che formano gli strati della rete neurale artificiale seguono un comportamento che imita quello dei neuroni biologici. In riferimento all'immagine seguente, il set $\{x_1, ..., x_n\}$ rappresenta l'input al neurone **j**, mentre $\{w_{1j}, ..., w_{nj}\}$ sono i pesi associati a ciascun input e determinano quanto ognuno di essi contribuisce all'output del neurone; la regolazione dei pesi permette al modello di apprendere. Il neurone esegue la somma $\sum_i^n (x_i \cdot w_{ij})$ e apporta un bias b_j che rappresenta il suo stato interno. Infine, applica una funzione f di attivazione per restituire l'output y_i.

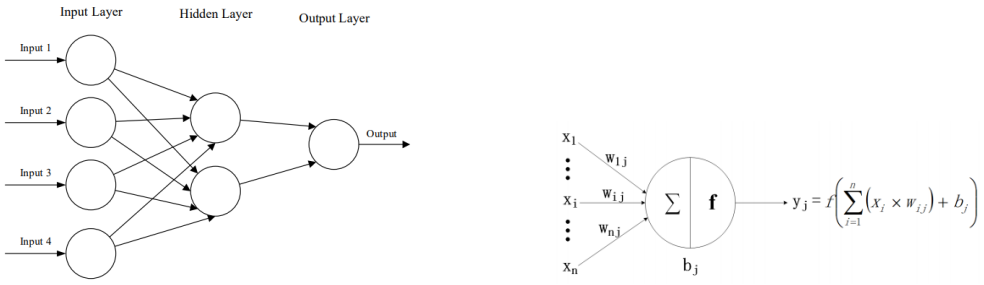

Figura 2.2. A sinistra: semplice rete neurale artificiale con tre strati. A destra: principio di funzionamento di un neurone artificiale.

Affinché la rete possa estrarre features complesse dai dati, deve poter sfruttare delle funzioni di attivazione non lineari; tra le più comuni ci sono:

- sigmoide $\sigma(x) = \frac{1}{1+e^{-x}}$: restituisce un output nel range $(0, 1)$, dunque risulta utile nel caso di classificazione binaria; tuttavia, ha problemi di vanishing gradient per input molto grandi o molto piccoli, con conseguente rallentamento del training.
- softmax $f(z_i) = \frac{e^{z_i}}{\sum_i^n e^{z_i}}$: si applica nel layer di output, restituisce un output nel range $(0, 1)$ ed è utilizzata per la classificazione multiclasse; trasforma in probabilità i logit z_i, che sono i valori grezzi iniziali per ciascuna classe restituiti dall'ultimo strato della rete.
- tangente iperbolica $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$: restituisce un output nel range $(-1, 1)$, dunque con valore medio pari a 0, e questo genera una sorta di normalizzazione

tale da velocizzare l'addestramento in alcune reti; tuttavia, anche questa funzione è soggetta al problema del vanishing gradient.

- ReLU $f(x) = max\{0, x\}$: la sua derivata può essere zero oppure uno e questo genera un effetto ottimale per la convergenza. Tuttavia, presenta il problema dei dying neurons, ovvero quando $x < 0$ il gradiente è nullo, dunque i neuroni non si aggiornano e non contribuiscono al modello.
- Leaky ReLU $f(x) = \begin{cases} x, & \text{se } x > 0, \\ \alpha x, & \text{altrimenti} \end{cases}$: risolve il problema dei neuroni spenti introducendo un piccolo gradiente anche per valori negativi, mediante α che assume valori piccoli.

Il calcolo del gradiente nelle reti neurali avviene tramite il metodo della backpropagation introdotto nel 1986, che prevede per ogni iterazione due fasi distinte. Prima di tutto si mantengono i pesi fissati e si propaga l'input attraverso la rete, dal primo all'ultimo strato. A questo punto si può confrontare l'output con il valore atteso, producendo un certo errore; questa informazione viene propagata a ritroso in tutti gli strati della rete permettendo a ogni neurone di contribuire all'aggiornamento dei pesi, in modo da ridurre l'errore nelle iterazioni successive.

2.2.1 Convolutional Neural Networks

Le Convolutional Neural Networks (CNNs) sono le reti più utilizzate nel campo del deep learning e permettono di eseguire compiti come, per esempio, il riconoscimento facciale, la guida autonoma e diverse diagnosi mediche. Infatti, le CNN sono un tipo di feedforward neural network in grado di estrarre caratteristiche dai dati, in particolare dalle immagini, mediante l'operazione matematica lineare di convoluzione.

Le reti neurali artificiali con layer completamente connessi sono in grado di analizzare le immagini, tuttavia non tengono conto della struttura spaziale delle immagini e trattano ogni pixel allo stesso modo. Le reti neurali convoluzionali non utilizzano layer completamente connessi; per esempio, un neurone nel primo hidden layer viene connesso soltanto con una regione localizzata dell'input layer, e tale regione è detta campo ricettivo locale del neurone. Questo riduce la complessità del modello rispetto a una rete neurale convenzionale. Il campo ricettivo, detto anche kernel, agisce come un filtro che scorre sull'input per rilevarne le caratteristiche. Grazie alla condivisione dei pesi, i neuroni dello stesso layer rilevano esattamente tutti la stessa caratteristica, ovvero applicano tutti lo stesso kernel, producendo una feature map; questo riduce il numero dei parametri e rende la rete invariante rispetto alle traslazioni, ovvero in grado di riconoscere alcune caratteristiche indipendentemente dalla loro posizione. Di solito, per non perdere informazioni sui bordi mentre il filtro scorre sull'immagine, è comune impostare anche un padding, che aumenta la dimensione dell'input aggiungendo zeri ai bordi dell'immagine.

L'operazione di convoluzione eseguita dal filtro è simile a quella descritta nella seconda immagine (2.2), dove però il prodotto $x_i \cdot w_{ij}$ è ora eseguito tra gli elementi della matrice di input e la matrice dei pesi del filtro, come illustrato nella figura seguente.

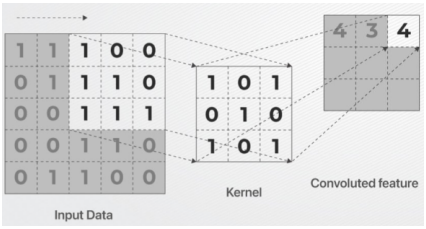

Figura 2.3. Operazione di convoluzione.

Il passo del filtro sull'immagine è modulato da un parametro chiamato stride, che definisce dunque la densità della convoluzione: maggiore è lo stride, minore è la densità della convoluzione, e viceversa.
Definite D la dimensione originale dell'input, p l'entità del padding, s l'entità dello stride e k la dimensione del filtro, la dimensione finale dell'output degli strati convoluzionali sarà: $D' = \frac{D-k+2p}{s} + 1$.

In genere, dopo gli strati convoluzionali, le CNN prevedono strati di pooling che semplificano l'informazione raccolta dalle convoluzioni condensando le feature map, e rafforzando la proprietà della rete di essere poco sensibile a piccole traslazioni dell'immagine. L'idea alla base del pooling è che una volta trovata una feature, la sua posizione esatta non è importante quanto la sua posizione rispetto ad altre features. Le due operazioni di pooling più utilizzate sono il max pooling, che seleziona il valore massimo all'interno di una regione, e l'average pooling, che invece seleziona il valore medio.

Lo strato finale di una CNN è uno strato completamente connesso che combina le caratteristiche estratte dagli strati precedenti per generare l'output voluto. A questo strato finale si applica una funzione di attivazione specifica in base al task che la rete deve adempiere.

2.3 Convolutional Autoencoder e Anomaly Detection

2.3.1 Autoencoder e Convolutional Autoencoder

L'autoencoder (AE) è un modello di deep learning basato sull'apprendimento non supervisionato, in grado di comprimere l'input per estrarne le caratteristiche più importanti e apprenderne una rappresentazione a bassa dimensionalità (detta latente), seguendo l'obiettivo di ricostruirlo lasciandone invariate le dimensioni e introducendo quante meno distorsioni possibili; quindi può autonomamente apprendere dai dati in input e può agire come metodo di riduzione dimensionale. La sua struttura è quella di una rete feedforward con tre layer totalmente connessi che formano le due parti caratteristiche dell'autoencoder convenzionale: encoder e decoder.

L'encoder si occupa della compressione degli input $x_i \in \mathbb{R}^n$ in una dimensione pari a $t < n$ numero di neuroni che formano l'hidden layer; lo scopo della compressione è l'estrazione delle features rilevanti dai dati, per utilizzarle come nuova rappresentazione dei dati. L'attivazione di ciascun neurone avviene tramite una

funzione di attivazione tra quelle già descritte e in questo modo si ottiene una codifica dell'input in un vettore di dimensioni ridotte $h_i \in \mathbb{R}^t$.
Il decoder, invece, si occupa di decomprimere le rappresentazioni h_i e decodificarle per ricostruire l'input originale con la stessa dimensione.
Il numero di neuroni nell'input layer e nell'output layer sono gli stessi per garantire che l'output abbia le stesse dimensioni dell'input, mentre il numero di neuroni nell'hidden layer assume un valore minore, affinché avvenga la compressione dei dati.

L'obiettivo del modello è minimizzare l'errore di ricostruzione ϵ tra l'input originale e l'output ricostruito, che può essere valutato con una delle metriche descritte nella sezione precedente.

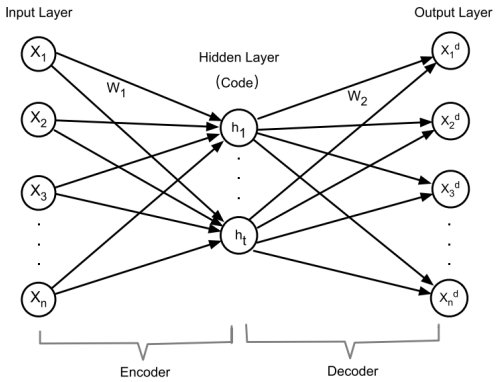

Figura 2.4. Semplice struttura di un autoencoder con tre strati.

Sostituendo gli strati completamente connessi con strati convoluzionali, si ottiene un Convolutional Autoencoder (CoAE); infatti, come visto, gli strati convoluzionali sono in grado di diminuire le features dei dati in input, mentre gli strati deconvoluzionali possono aumentarle, riproducendo così il principio di funzionamento di un autoencoder convenzionale, con il vantaggio di essere particolarmente adatto per l'analisi delle immagini.

2.3.2 CoAE per Anomaly Detection

Uno degli utilizzi più comuni per i modelli CoAE è l'anomaly detection.
Si definiscono anomalie o eventi rari quei dati che non sono conformi a una nozione di comportamento normale per un dato modello; il rilevamento delle anomalie, quindi, ha come obiettivo trovare quei dati che, date le osservazioni precedenti, non aderiscono alle norme previste. L'ipotesi di partenza è che i campioni anomali rappresentino una percentuale molto piccola dell'intero dataset; dunque, anche quando sono disponibili dati etichettati, i dati normali sono più facilmente disponibili.

Un primo approccio potrebbe essere quello di apprendimento supervisionato, ovvero un apprendimento nel quale si hanno a disposizione dati normali etichettati e dati anomali etichettati per tutti i tipi di anomalie che potrebbero verificarsi. Questo approccio riformula così il problema del rilevamento delle anomalie in un problema

di classificazione; ma non sempre è applicabile, poiché in alcuni casi le anomalie assumono molte forme diverse ed è impossibile etichettarle tutte. Pertanto, sono preferibili approcci di apprendimento non supervisionato che apprendono direttamente dai dati in input. senza avere dati anomali etichettati. La strategia di base è quella di apprendere prima il comportamento normale dei dati, e poi sfruttare questa conoscenza per identificare le anomalie.

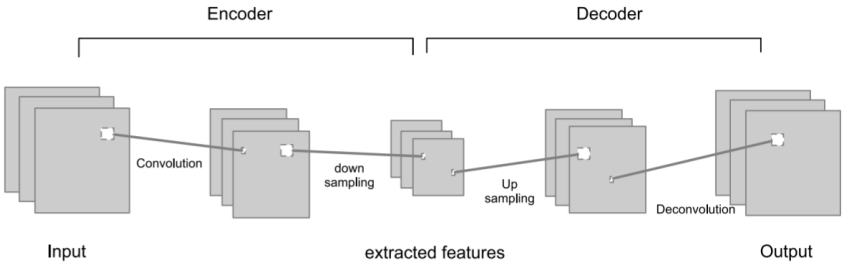

Figura 2.5. Visualizzazione grafica di un CoAE.

Sfruttando le capacità dei CoAE, quando si ricostruiscono eventi anomali si registra in genere un errore di ricostruzione più grande rispetto a quello sui dati normali; questo errore ϵ si può utilizzare come punteggio di anomalia e si può impostare una soglia θ tale per cui se $\epsilon < \theta$ allora il dato è definito normale, altrimenti è definito anomalo.

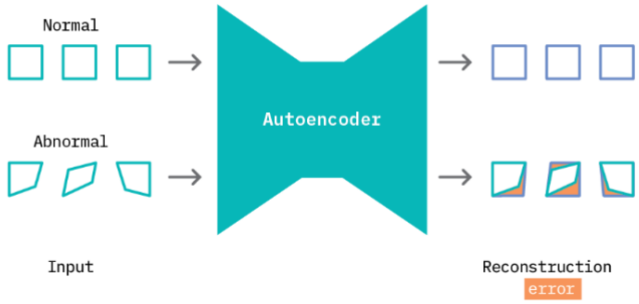

Figura 2.6. Schema riassuntivo del funzionamento di un autoencoder per anomaly detection.

La rilevazione delle anomalie è diventata parte integrante della moderna ricerca sulla fisica delle particelle e permette di identificare eventi rari e inaspettati che potrebbero essere indicativi di nuovi fenomeni o di fisica BSM. Siccome i dati di fisica delle particelle non sempre possono essere etichettati a priori, a causa della presenza di eventi di fondo oppure sovrapposizioni quantistiche tra segnale e fondo, un approccio che classifica in base alle etichette o che prevede una certa etichetta minimizzando una funzione di costo non è sempre attuabile. Inoltre, l'approccio tradizionale per la ricerca di nuovi segnali comporta la ricerca di un segnale specifico e la massimizzazione della sensibilità dell'analisi per quel singolo segnale. Questo approccio non è utile per la ricerca di nuovi modelli fisici; così, per poter rilevare simultaneamente diversi nuovi scenari, si implementano modelli di anomaly detection.

Capitolo 3

Applicazione dell'anomaly detection nell'identificazione dei jet adronici

In questo capitolo viene presentata una possibile applicazione di un modello autoencoder convoluzionale che, mediante anomaly detection, identifica jet adronici associati a quark top e bosoni W e Z che decadono con un grande boost di Lorentz.

A tal fine, si utilizza un dataset composto da 30000 immagini, ciascuna rappresentante un jet adronico. Queste immagini sono istogrammi bidimensionali che mostrano la distribuzione del momento trasverso p_T depositato in ogni bin di una griglia 100×100 centrata sull'asse del jet e sita nel piano (η, ϕ), come si può vedere nella figura seguente.

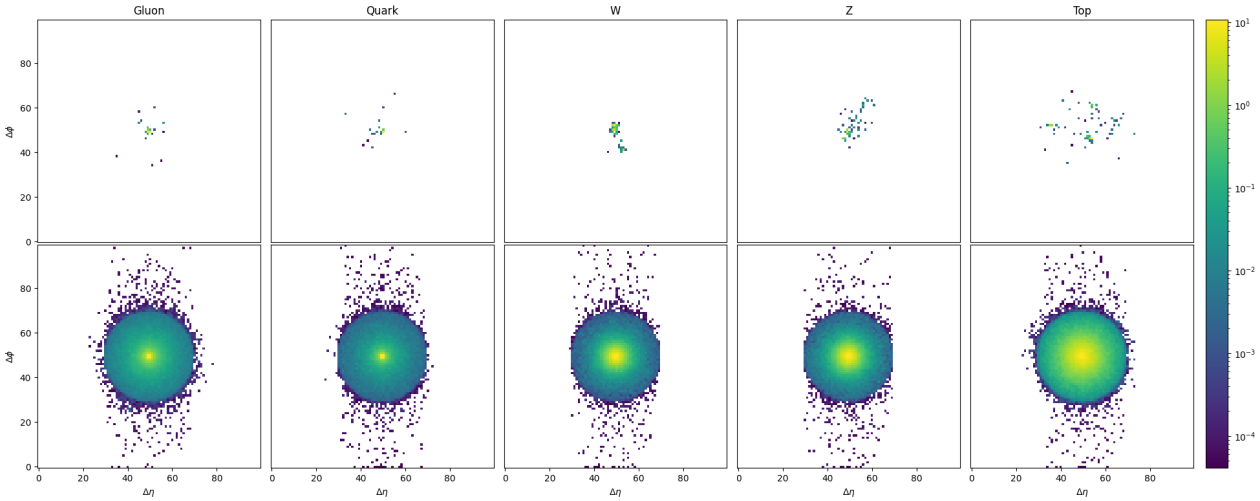

Figura 3.1. Panoramica del dataset. In alto: singola immagine per ciascuna classe. In basso: somma di tutte le immagini per ciascuna classe. Le immagini sono rappresentate con una normalizzazione logaritmica.

3.1 Preprocessing del dataset

Il dataset contiene esempi per cinque classi diverse: gluone, quark, bosone W, bosone Z e quark top; la suddivisione è tale da rendere il dataset piuttosto bilanciato, con seimila esempi circa per ciascuna classe. Per facilitare l'analisi successiva, è stato necessario normalizzare le immagini, facendo in modo che ogni pixel assuma un valore compreso tra $[0, 1]$ senza alterarne l'informazione fisica.
Gli esempi contenuti nel dataset sono stati successivamente divisi tra eventi normali ed eventi anomali. In questo caso, gli eventi normali sono rappresentati dai jet adronici prodotti da quark e gluoni; invece, gli eventi anomali che il modello deve individuare sono rappresentati dai jet adronici prodotti dai bosoni W e Z e dal quark top. Con i dati a disposizione, dopo la suddivisione, risultano 11994 eventi normali e 18006 eventi anomali.
Nei dati normali è stata eseguita un'ulteriore suddivisione per generare tre dataset: il training set, contenente il 80% degli eventi normali (9595 immagini), il validation set, contenente il 10% dei restanti eventi (1199 immagini) e il test set contenente l'altro 10% degli eventi restanti (1200 immagini).

3.2 Definizione del modello CoAE

Il modello autoencoder convoluzionale utilizzato è composto da un encoder e un decoder con la seguente architettura:

- Encoder: quattro layer convoluzionali con attivazione ReLU che applicano ciascuno un kernel 4×4 con uno stride di 2 e un padding di 1. Ogni strato convoluzionale dimezza le dimensioni dell'input e ne cattura alcune informazioni globali, cosicché l'input iniziale 100×100 è compresso infine in un'immagine 6×6. Ogni strato, inoltre, aumenta il numero dei filtri applicati seguendo le potenze di due, da 32 filtri nello strato iniziale a 256 nell'ultimo. Dopo ogni strato convoluzionale viene applicata una batch normalization, che ha lo scopo di normalizzare l'output e rendere più stabile l'allenamento della rete neurale. Infine, l'output degli strati convoluzionali viene passato ad un layer completamente connesso che lo comprime ulteriormente in un vettore di dimensione 5, corrispondente alla dimensione scelta per lo spazio latente.
- Decoder: quattro layer convoluzionali trasposti con attivazione ReLU che decomprimono l'output dell'encoder. Per prima cosa, il vettore latente viene proiettato in uno spazio più grande di dimensione 9216 ($256 \times 6 \times 6$). Successivamente, i quattro layer deconvoluzionali applicano ciascuno un kernel 4×4 con uno stride di 2 e un padding di 1 per ricostruire l"input originale di dimensioni 100×100. Anche in questo caso vengono inseriti degli strati di batch normalization per stabilizzare la ricostruzione. Infine, lo strato che riproduce l'output finale ha come funzione di attivazione la sigmoide, che restituisce un output normalizzato adatto all'utilizzo con la BCE come funzione obiettivo.

3.2.1 Addestramento del modello

Il modello così definito è stato addestrato con l'obiettivo di apprendere una rappresentazione latente dei dati normali di training e di ricostruire tali immagini. Le performance del modello sono state misurate mediante la Binary Cross-Entropy, poiché tale funzione è progettata per lavorare con dati normalizzati e valuta la differenza tra pixel previsti e pixel veri in maniera più severa rispetto alla MSE o alla MAE. Come ottimizzatore è stato utilizzato Adam con un learning rate iniziale pari a 0.001, regolato da uno scheduler che ne riduce il valore quando le performance del modello non migliorano per cinque epoche consecutive. Per epoca si intende un passaggio completo attraverso tutto il dataset di addestramento. L'ottimizzazione della funzione di costo si è svolta per ventotto epoche, monitorata da un early stopping, una tecnica che interrompe il training quando la funzione di perdita non mostra miglioramenti significativi per un numero predefinito di epoche consecutive, evitando così l'overfitting.

Il modello così addestrato è stato testato sul test set, valutando in particolare la capacità del modello di ricostruire immagini normali mai viste in fase di addestramento; sulla base di tali ricostruzioni è stato possibile ricavare una distribuzione dell'errore di ricostruzione sul test set, ottenendo una distribuzione con valore massimo pari a $(1.327 \pm 0.043) \cdot 10^{-3}$ e un valore medio $(1.809 \pm 0.023) \cdot 10^{-3}$.

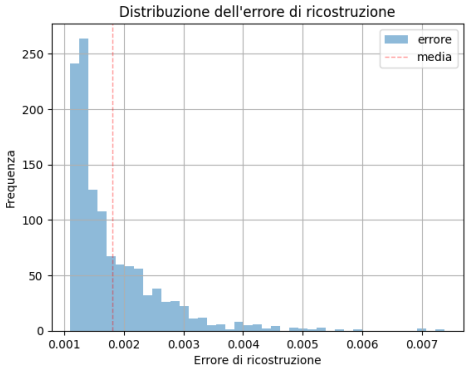

Figura 3.2. Distribuzione dell'errore di ricostruzioni per le immagini normali del test set.

3.3 Anomaly detection

Per testare la capacità del modello di riconoscere immagini anomale, ovvero non appartenenti alle classi di immagini con le quali è stato addestrato, si può sfruttare l'errore di ricostruzione di tali immagini. In particolare, confrontando le distribuzioni dell'errore di ricostruzione per i dati normali e quelli anomali, si osserva che queste risultano generalmente separate, con errori di ricostruzione generalmente maggiori per le anomalie. Ciò consente di definire una soglia di decisione ottimale per distinguere i due gruppi e identificare le immagini anomale.

3.3.1 Metriche utilizzate

Per valutare le performance del modello sull'anomaly detection sono state utilizzate alcune metriche per le quali è opportuno definire le quantità chiave che ne permettono la stima. Si definiscono quindi true positive TP gli eventi anomali riconosciuti correttamente come anomalie, mentre i false positive FP sono gli eventi normali erroneamente classificati come anomalie; allo stesso modo, i true negative TN sono gli eventi normali classificati correttamente come normali, mentre i false negative FN indicano gli eventi anomali che il modello non è riuscito a classificare come anomalie. Così è possibile definire le metriche:

- precisione: $\frac{TP}{TP+FP}$, fornisce una stima di quanto siano accurate le predizioni positive.
- sensibilità (recall): $\frac{TP}{TP+FN}$, fornisce una stima della capacità del modello di identificare tutti i veri positivi.
- F1-score: $\frac{2TP}{2TP+FP+FN}$, media armonica tra precisione e sensibilità, utile nel caso di classi sbilanciate.
- ROC: Receiver Operating Characteristic, rappresenta su un piano il tasso di veri positivi ($TPR = \frac{TP}{TP+FN}$) in funzione del tasso di falsi negativi ($FPR = \frac{FP}{FP+TN}$) al variare della soglia di decisione.
- AUC: Area Under ROC, area sottesa alla curva ROC, compresa nell'intervallo $[0.5, 1.0]$ che va rispettivamente da un classificatore casuale a un classificatore ideale; fornisce una misura aggregata delle prestazioni del modello su tutte le possibili soglie di decisione.

La soglia di decisione ottimale è stata scelta tramite la curva ROC, in particolare individuando il punto sulla curva che minimizza la distanza euclidea dal punto $(0, 1)$, ovvero la situazione in cui non ci sono falsi positivi e tutte le anomalie sono correttamente classificate.
Le incertezze su ogni metrica sono state valutate attraverso bootstrap, una tecnica che utilizza campionamenti ripetuti con sostituzione dai dati per avere più stime della statistica di interesse.

3.3.2 Caso del quark top

Nel caso del dataset contenente esempi relativi al quark top, si ottiene una distribuzione dell'errore di ricostruzione con un massimo in $(2.201 \pm 0.072) \cdot 10^{-3}$ e un valore medio di $(2.816 \pm 0.012) \cdot 10^{-3}$.

AUC	Soglia	Precisione	Sensibilità	F1-score
0.839 ± 0.007	$(2.014 \pm 0.051) \cdot 10^{-3}$	0.938 ± 0.003	0.824 ± 0.023	0.877 ± 0.013

Tabella 3.1. Metriche con relative incertezze dell'anomaly detection sulle immagini da quark top.

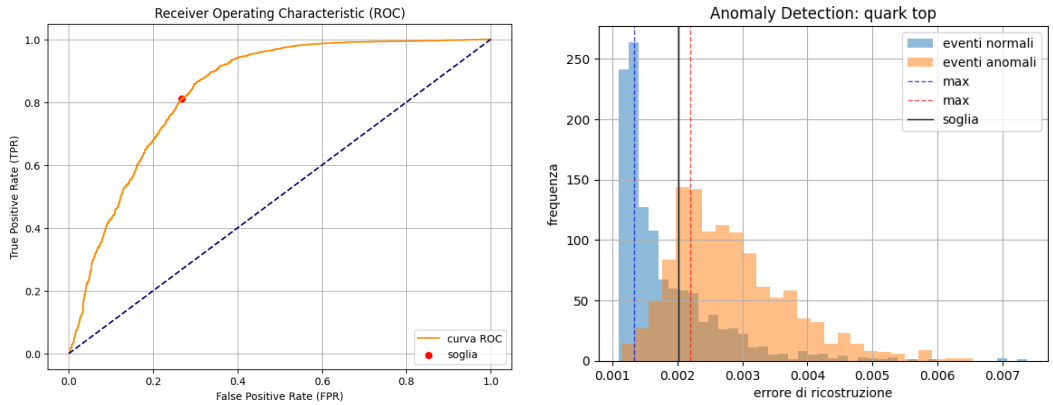

Figura 3.3. A sinistra: curva ROC con soglia ottimale. A destra: soglia di decisione sulle distribuzioni dell'errore di ricostruzione.

3.3.3 Caso del bosone W

Nel caso del dataset contenente esempi relativi al bosone W, si ottiene una distribuzione dell'errore di ricostruzione con un massimo in $(1.774 \pm 0.037) \cdot 10^{-3}$ e un valore medio di $(1.889 \pm 0.059) \cdot 10^{-3}$.

AUC	Soglia	Precisione	Sensibilità	F1-score
0.621 ± 0.009	$(1.638 \pm 0.025) \cdot 10^{-3}$	0.892 ± 0.004	0.687 ± 0.019	0.776 ± 0.013

Tabella 3.2. Metriche con relative incertezze dell'anomaly detection sulle immagini da bosone W.

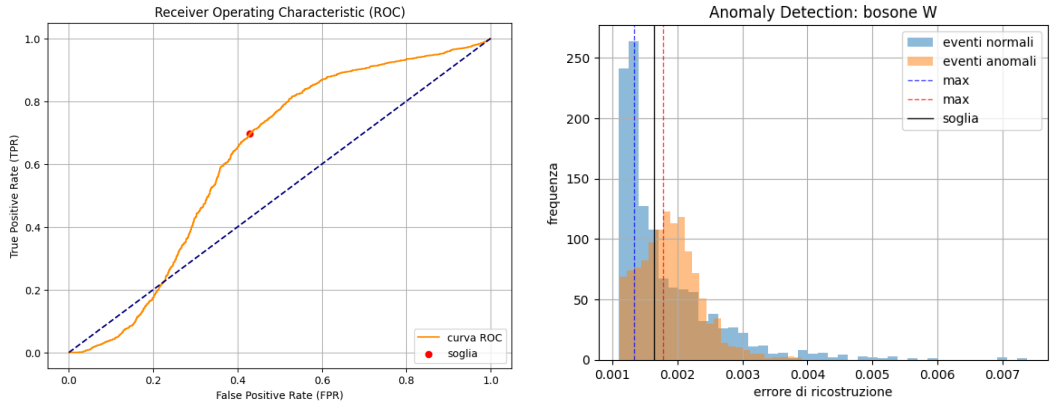

Figura 3.4. A sinistra: curva ROC con soglia ottimale. A destra: soglia di decisione sulle distribuzioni dell'errore di ricostruzione.

3.3.4 Caso del bosone Z

Nel caso del dataset contenente esempi relativi al bosone W, si ottiene una distribuzione dell'errore di ricostruzione con un massimo in $(1.971 \pm 0.037) \cdot 10^{-3}$ e un valore medio di $(1.970 \pm 0.064) \cdot 10^{-3}$.

AUC	Soglia	Precisione	Sensibilità	F1-score
0.655 ± 0.010	$(1.667 \pm 0.015) \cdot 10^{-3}$	0.899 ± 0.004	0.723 ± 0.013	0.802 ± 0.008

Tabella 3.3. Metriche con relative incertezze dell'anomaly detection sulle immagini da bosone Z.

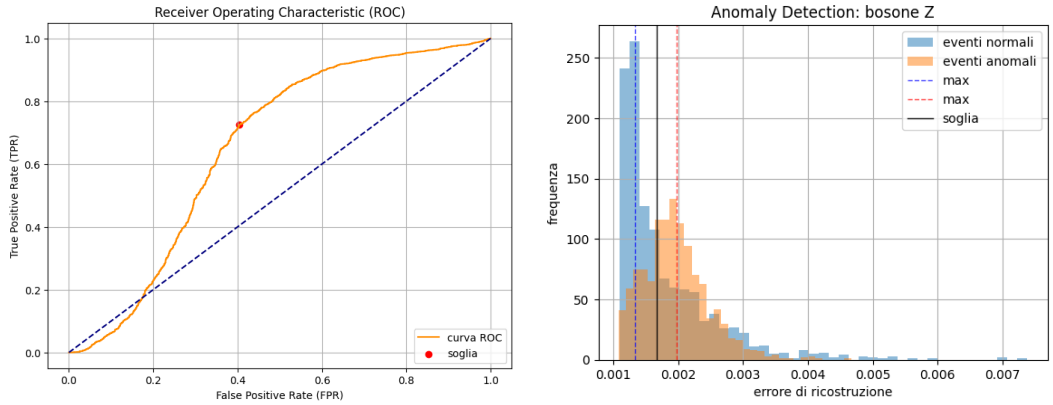

Figura 3.5. A sinistra: curva ROC con soglia ottimale. A destra: soglia di decisione sulle distribuzioni dell'errore di ricostruzione.

3.4 Conclusioni

In conclusione, i risultati ottenuti evidenziano che il modello è in grado di distinguere efficacemente i jet adronici prodotti dal decadimento di quark top con un grande boost di Lorentz. Questo è evidente dall'analisi delle metriche, in particolare dall'AUC, che mostra prestazioni complessivamente migliori rispetto al caso dei jet prodotti dal decadimento di bosoni W e Z altamente energetici.
Le differenze nelle prestazioni possono essere attribuite alle caratteristiche delle immagini considerate come normali durante l'addestramento del modello; infatti, le immagini normali rappresentano jet prodotti dal decadimento di quark e gluoni e sono caratterizzate dalla presenza di un solo jet adronico. Nel caso del decadimento del quark top, le immagini presentano una sottostruttura più complessa, composta da tre jet adronici, e questa caratteristica facilita l'identificazione di tali eventi come anomalie. Al contrario, i jet adronici derivanti dal decadimento di bosoni W e Z hanno una sottostruttura formata da due jet adronici, e questo rende più difficile la loro classificazione come anomalie.

Bibliografia

[1] Roy Schimmel Brener. *Estimating the Sensitivity to New Resonances Decaying to Boosted Quark Pairs and Produced in Association with a Photon in the ATLAS Experiment*. URL: https://cds.cern.ch/record/2753546/files/CERN-THESIS-2020-291.pdf.

[2] Jochen Jens Heinrich. *Reconstruction of boosted $W^{\pm}$ and Z^{0} bosons from fat jets*. URL: https://cds.cern.ch/record/1956424/files/CERN-THESIS-2014-152.pdf.

[3] Andrew J. Larkoski, Ian Moult e Benjamin Nachman. *Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning*. URL: https://www.sciencedirect.com/science/article/pii/S0370157319303643?via%3Dihub.

[4] *Jets and Jet Substructure at Future Colliders*. URL: https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2022.897719/full.

[5] *Boosted objects: a probe of beyond the standard model physics*. URL: https://www.slac.stanford.edu/pubs/slacpubs/15000/slac-pub-15081.pdf.

[6] CERN. *The Large Hadron Collider*. URL: https://home.cern/science/accelerators/large-hadron-collider.

[7] CERN. *LHC the guide FAQ*. URL: https://www.home.cern/resources/brochure/knowledge-sharing/lhc-facts-and-figures.

[8] CERN. *The ATLAS Detector*. URL: https://atlas.cern/Discover/Detector.

[9] Eric M. Metodiev. *The Fractal Lives of Jets*. URL: https://www.ericmetodiev.com/post/jetformation/.

[10] Pauline Gagnon. *The Standard Model: a beautiful but flawed theory*. URL: https://www.quantumdiaries.org/2014/03/14/the-standard-model-a-beautiful-but-flawed-theory/.

[11] CERN. *The Standard Model*. URL: https://www.home.cern/science/physics/standard-model.

[12] Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep Learning*. URL: https://www.deeplearningbook.org/front_matter.pdf.

[13] Pengzhi Li, Yan Pei e Jianqiang Li. *A comprehensive survey on design and application of autoencoder in deep learning*. URL: https://www.sciencedirect.com/science/article/pii/S1568494623001941?casa_token=7jQgvbANa_0AAAAA:n1zJnL28TOoCu4PH8JUV7PoLClWfMl8Qz-s5FuLCoMGEGYsADSTCsJQDn3bp6Dl0uAeT9hwb

[14] Zewen Li et al. *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects*. URL: https://ieeexplore.ieee.org/document/9451544.

[15] Michael Nielsen. *Neural Networks and Deep Learning*. URL: http://neuralnetworksanddeeplearning.com/index.html.

[16] Yifei Zhang. *A Better Autoencoder for Image: Convolutional Autoencoder*. URL: https://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf.

[17] Zhaomin Chen et al. *Autoencoder-based network anomaly detection*. URL: https://ieeexplore.ieee.org/document/8363930/authors.

[18] Cloudera Fast Forward Labs. *Deep Learning for Anomaly Detection*. URL: https://ff12.fastforwardlabs.com/.

[19] Vasilis Belis, Patrick Odagiu e Thea Klaeboe Aarrestad. *Machine learning for anomaly detection in particle physics*. URL: https://www.sciencedirect.com/science/article/pii/S2405428324000017.

Ringraziamenti

In questa tesi, grazie a una preziosa e costante guida, ho potuto unire la mia passione per la fisica delle particelle all'intelligenza artificiale. Scoprire la sinergia tra questi due campi in continua evoluzione, ha reso la stesura di questa dissertazione piacevole ed entusiasmante in ogni suo momento.

Questo lavoro rappresenta il culmine di un percorso iniziato molti anni fa, quando, l'estate prima di cominciare il liceo, trovai un certo libro di fisica che cominciai a sfogliare per scoprire di cosa trattasse questa materia a me ancora sconosciuta, e che presto avrei trovato ad attendermi sui banchi di scuola. A guidarmi era una forte curiosità che non sapevo spiegarmi, e che divenne subito interesse non appena compresi il valore di cosa avevo davanti. Scoprire le semplici e affascinanti relazioni tra velocità, spazio e tempo accese quella parte della mia curiosità che ha reso questa disciplina il centro del mio interesse.

Il percorso che mi ha portato qui non è stato facile, tantomeno lineare. Ci sono stati tanti dubbi, ostacoli ed errori. Ma non sono mancate le soddisfazioni, che ogni volta davano senso a tutti gli sforzi fatti. Quindi, alla fine di questo lungo viaggio, guardando indietro, posso dire che per me è valsa la pena di provarci e riuscire lì dove si sono arresi gli altri.

Adesso, caro lettore, mi rivolgo a te. Se stai leggendo queste righe è perché hai apprezzato le idee esposte in questa tesi e l'hai letta fin qui, o forse perché hai scelto di saltare direttamente a questa pagina finale. Qualunque sia il motivo, apprezzo prima di tutto la tua curiosità, e ci tengo a dirti che sei una delle persone che desidero ringraziare sinceramente. La tua vicinanza e il tuo supporto, in qualche modo, sono stati fondamentali per rendere possibile tutto questo. Hai condiviso con me i successi, ma soprattutto i dubbi e i fallimenti che mi hanno permesso di crescere e di arrivare qui.
Se oggi queste parole giungono ai tuoi occhi, è grazie al contributo che, direttamente o indirettamente, mi hai offerto lungo il percorso. Questo traguardo, quindi, non è soltanto personale, ma è anche il riflesso di tutte le esperienze che abbiamo condiviso.

Con gratitudine, Gabriel.