UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Inverse Beta Decay events selection in JUNO using Machine Learning algorithms

Sviluppo di tecniche di Machine Learning per la selezione di interazioni di neutrini da reattori nell'esperimento JUNO

Relatore

Prof. Alberto Garfagnini

Correlatore

Dott. Andrea Serafini

Laureando

Fabio Cufino

Introduzione

- **JUNO detector**
- **JUNO signal and background**

JUNO detector

Il Jiangmen Underground Neutrino Observatory (JUNO) è in fase di realizzazione, sotto la collina Dashi (Sud della Cina).
Obiettivo principale: rilevazione degli $\overline{\nu}_e$

Central Detector (CD):

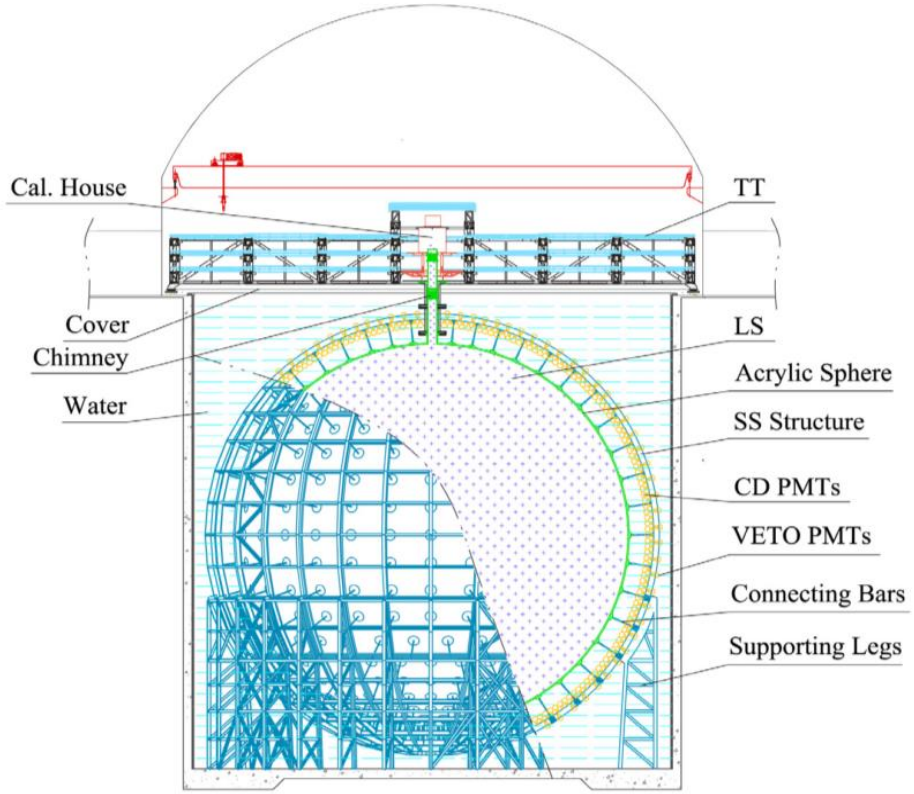
- rilevatore compatto
- risoluzione energetica effettiva di $\sigma_E/E = 3\%/\sqrt{E(MeV)}$
- posto all'interno di una vasca sferica di acrilico immersa in una piscina d'acqua: **Rivelatore Cherenkov**

Liquido Scintillatore (LS):

- 20 kton
- Linear Alkyl-Benzene, con una densità di 0.859 g/mL.

Top Tracker (TT):

- un insieme di un array di scintillatori plastici, disposti per misurare le tracce dei muoni.



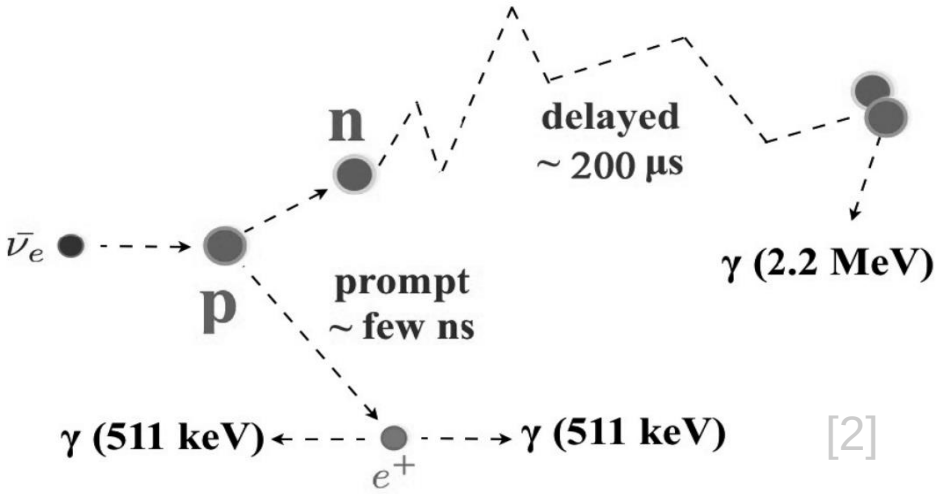
Introduzione

- JUNO detector
- JUNO signal and background

JUNO signal and background

Gli antineutrini del segnale sono prodotti dalla fissione nucleare dalle centrali di Taishan e Yangjiang, a circa 52.5 km dal rilevatore. (57.4 ev/day)

Inverse Beta Decay (IBD): $\overline{\nu}_e + p \rightarrow e^+ + n$



[2]

1. Prompt Signal

- Il e^+ deposita l'energia nel liquido scintillatore attraverso un processo di ionizzazione.

2. Delayed Signal

Dopo un ritardo medio di circa 200 μs

- Il n viene catturato dall'idrogeno (~99%) → Singolo fotone da 2.2 MeV
- Il n viene catturato sul carbonio (~1%) → Segnale con un'energia totale di 4.9 MeV

JUNO signal and background

➢ L'esperimento JUNO è progettato per minimizzare il background da varie fonti, ma alcuni segnali di fondo sono inevitabilmente prodotti nel rivelatore

Accidental Background

(~ 130 000 ev/day)

- **Definizione**: coincidenza di due eventi non correlati.
- **Origine**: Decadimenti radioattivi di isotopi (es. 238U, 232Th, 40K)
- **Effetti**: Simulano il segnale IBD

<u>Prompt</u>
Decadimenti β

<u>Delayed</u>
Decadimenti γ
Produzione di neutroni

Correlated Background

(~ 4 ev/day)

- **Definizione**: unica interazione fisica
- **Origine**: Background cosmogenici, geoneutrini, neutrini atmosferici, antineutrini provenienti dai reattori nucleari nel mondo
- **Effetti**: Il segnale è indistinguibile da IBD

→ Riduzione significativa possibile solo su eventi Accidental Background
→ **Obiettivo della tesi**: Studio di strategie per questa riduzione

Analisi

- **Feature Table**
- **Modelli**

Feature Table

- Due set di dati distinti
- Simulazioni Monte Carlo condotte tramite il software SNiPER.

1. ***IBD dataset***: potenziali eventi IBD, supponendo che le fonti di antineutrini siano i reattori Taishan e Yangjiang.

Per ogni evento:

SimID	(x,y,z)	E	t
Una coppia prompt-delayed originata da un evento IBD ha stesso SimID.	coordinate del punto all'interno del rivelatore dove è avvenuto l'evento	l'energia dell'evento	l'istante temporale in cui è avvenuto l'evento.

2. ***BKG dataset***: eventi di radioattività (*Accidental Background*)
   - La struttura del file è identica al dataset IBD
   - Natura non correlata degli eventi → SimID unico per ogni evento

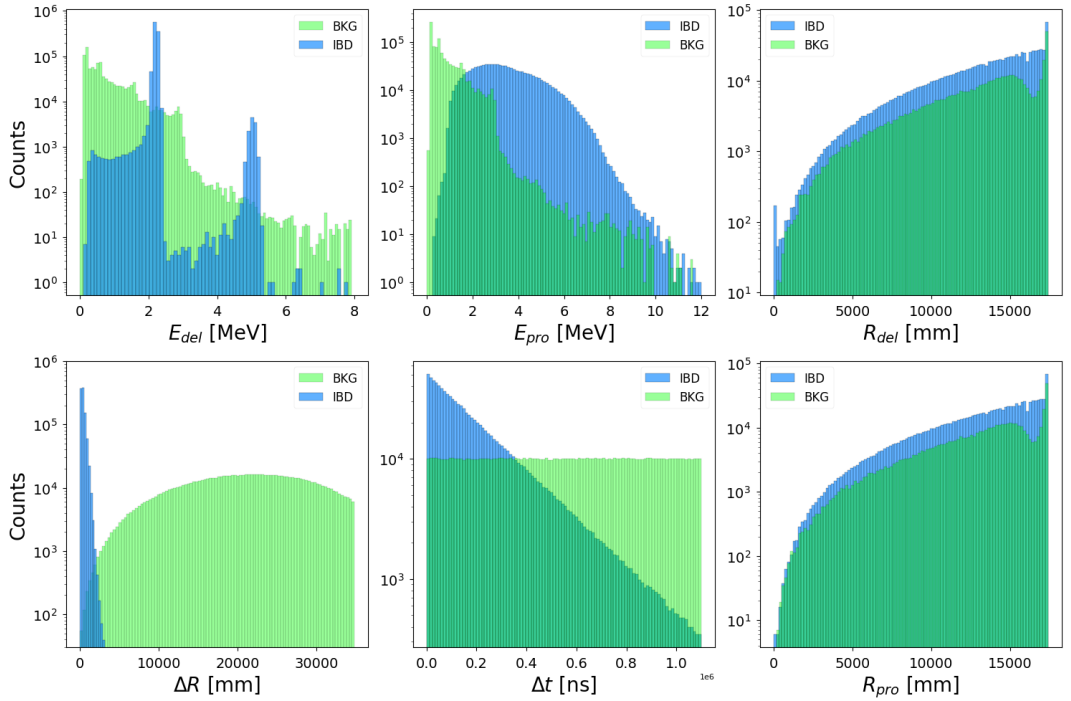
- Unione dei due dataset

Feature Table

- *Feature Table*: tutte le possibili coppie di eventi all'interno del dataset unificato, senza ripetizioni
- Codice sviluppato in Python, parallelizzato tramite Numba
- CloudVeneto, 14 CPU core

Per ogni coppia *prompt-delayed:*

$\mathbf{E_{del}}$ $\mathbf{E_{pro}}$	Energia del delayed e propmt signal
$\mathbf{R_{del}}$ $\mathbf{R_{pro}}$	Distanza del delayed e propmt signal
$\mathbf{\Delta t}$	Differenza temporale tra due eventi
$\mathbf{\Delta R}$	La differenza spaziale tra due eventi



Labelling:

Label 1: Coppia di eventi con stesso SimID

Label 0: Coppia di eventi non correlati con SimID differente

Analisi

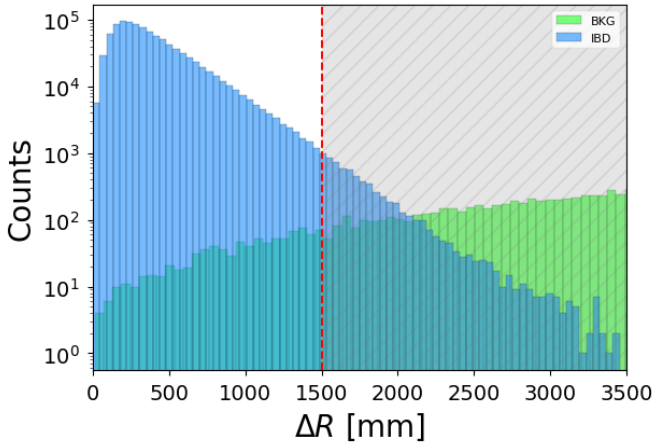
- Feature Table
- **Modelli**

Manual Cut

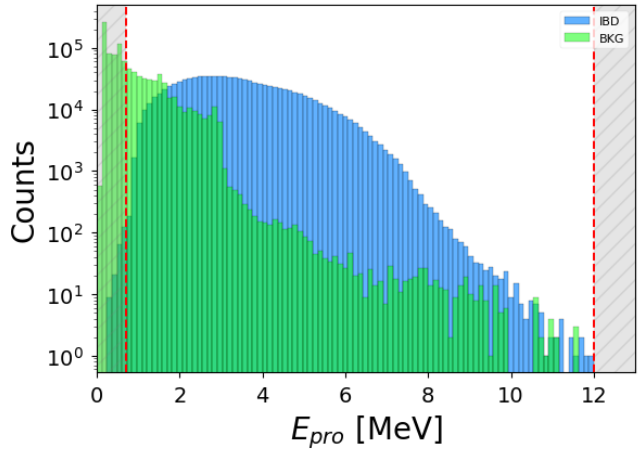
Lo stato dell'arte della classificazione al momento

L'algoritmo si basa su criteri di selezione applicati alla feature table: "**tagli**", adottati seguendo le indicazioni della letteratura [1]

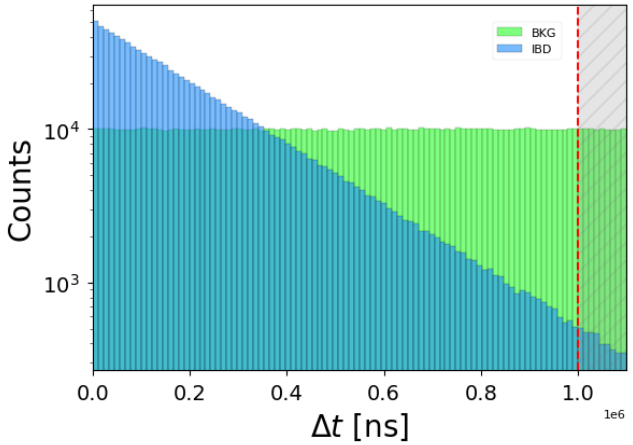
$\Delta R < 1500mm$



$0.7MeV < E_{pro} < 12MeV$

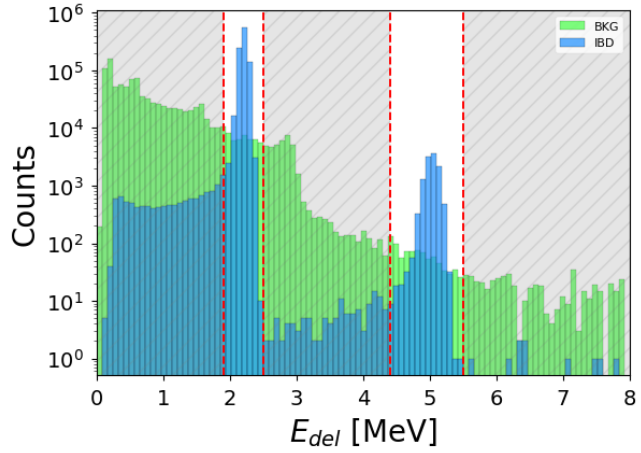


$\Delta t < 1ms$



$1.9MeV < E_{del} < 2.5MeV$

$4.4MeV < E_{del} < 5.5MeV$

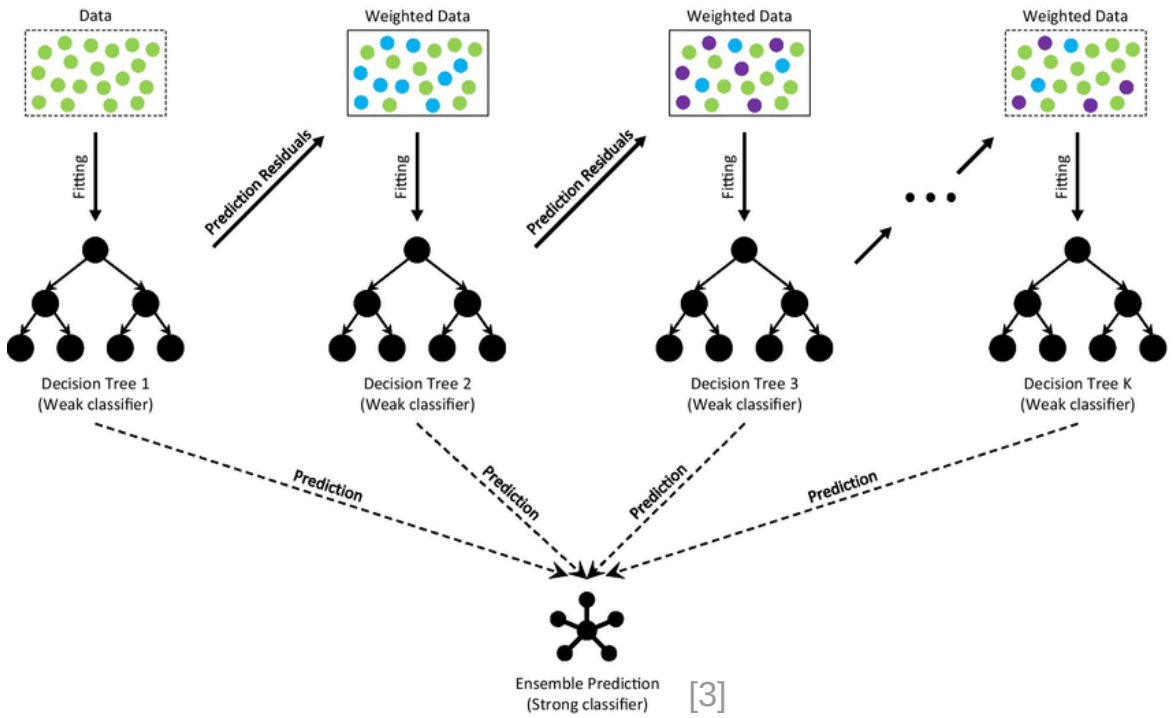


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Machine Learning

- **Boosted Decision Trees**

Boosted Decision Trees

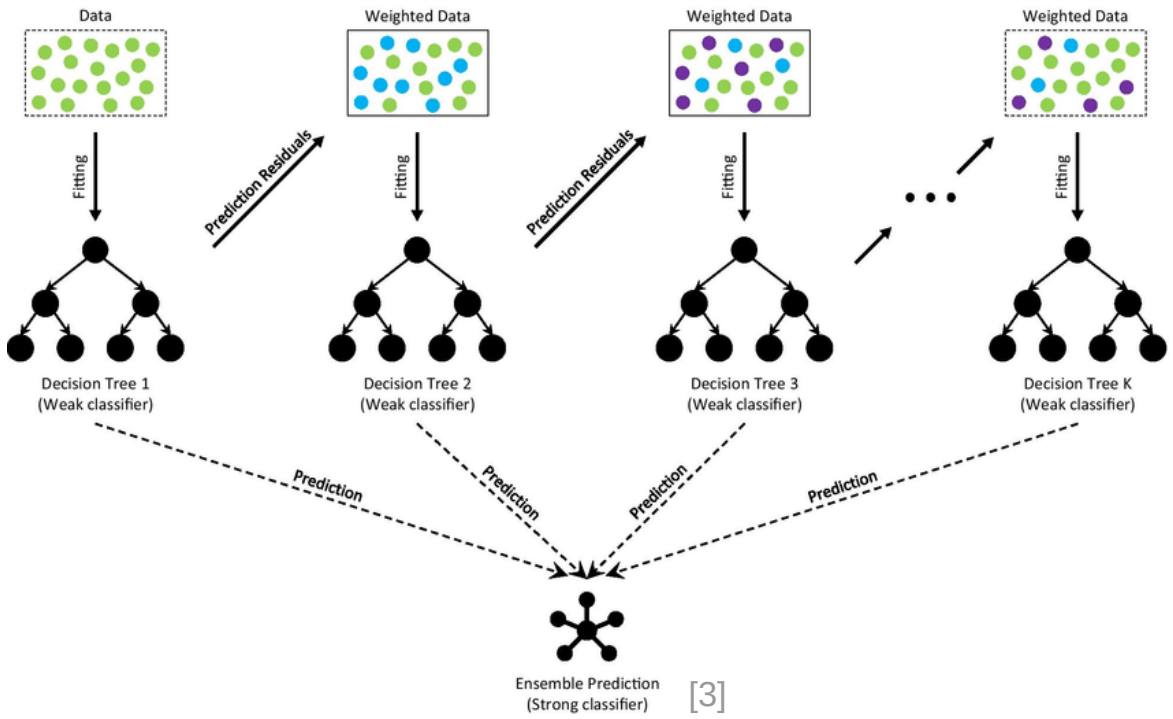


[3]

Introduzione: metodo avanzato di Supervised Learning che combina i punti di forza di due tecniche: gli alberi decisionali e il boosting.

Struttura: Serie di alberi decisionali (weak classifier) creati iterativamente.
La previsione finale dell'ensemble (Strong Classifier) si ottiene sommando le previsioni di ciascun albero, con un peso che dipende dall'efficacia di quell'albero.

Boosted Decision Trees



[3]

Processo di Apprendimento

Ogni nuovo albero

- addestrato sui "residui" del modello precedete: gli errori tra le previsioni del modello corrente e i valori reali.
- progettato per "imparare" da questi residui, ossia per correggere gli errori commessi dagli alberi precedenti. (**Boosting)**

- XGBoost è un'implementazione di questo metodo: l'elaborazione parallela

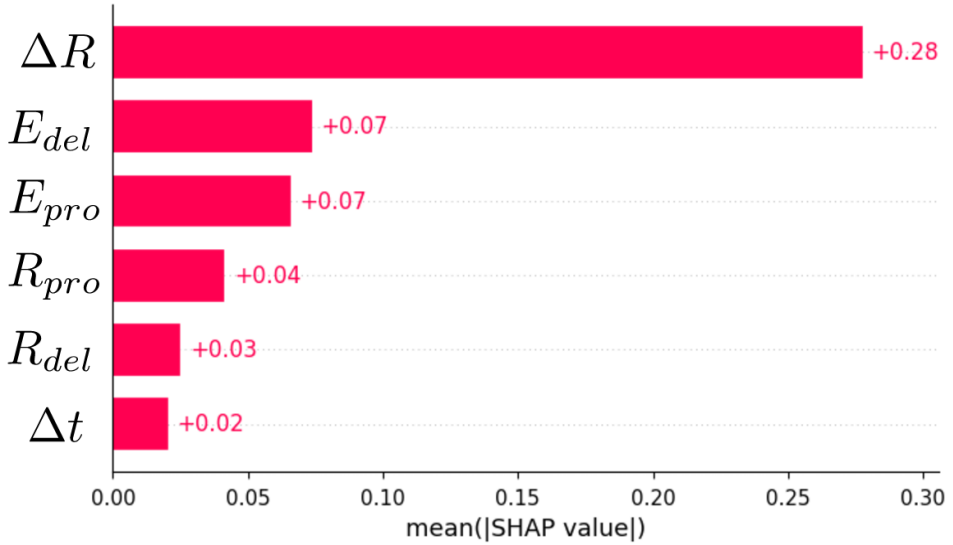
BDT: interpretazione

Interpretazione del modello

Libreria **SHAP** (SHapley Additive exPlanations)

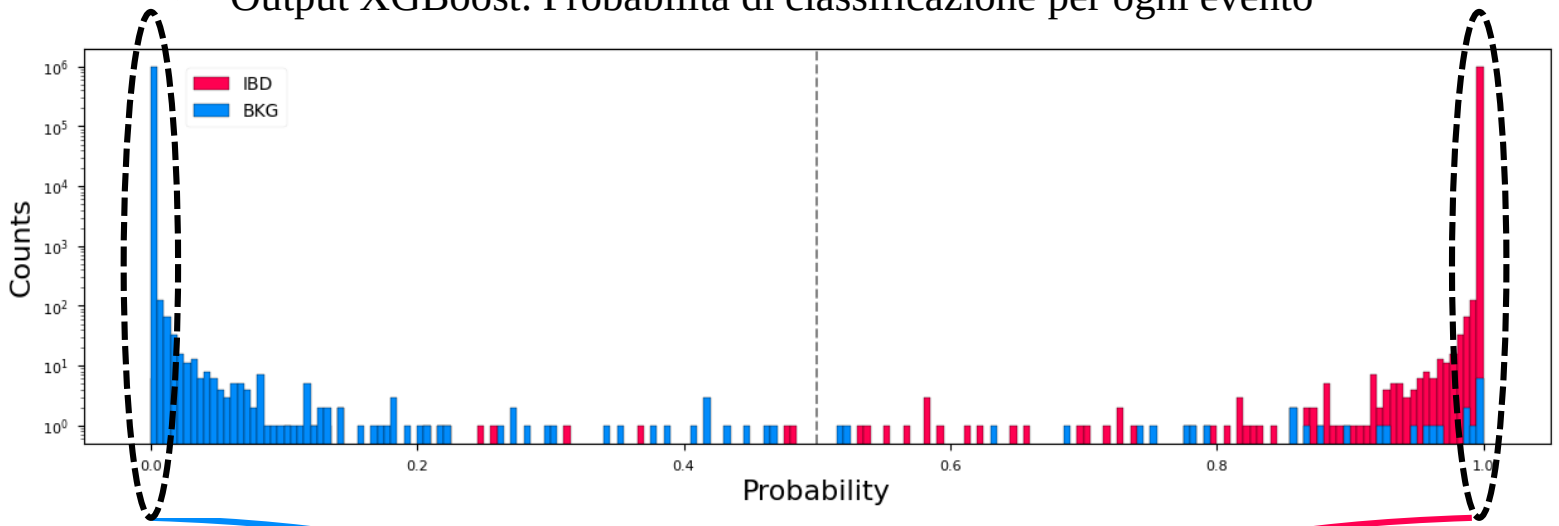
SHAP value: misura del contributo medio per una specifica feature alle previsioni del modello, tenendo conto di tutte le combinazioni possibili in cui questa feature può interagire con le altre.

→ Misura l'importanza della feature per il modello



BDT: interpretazione

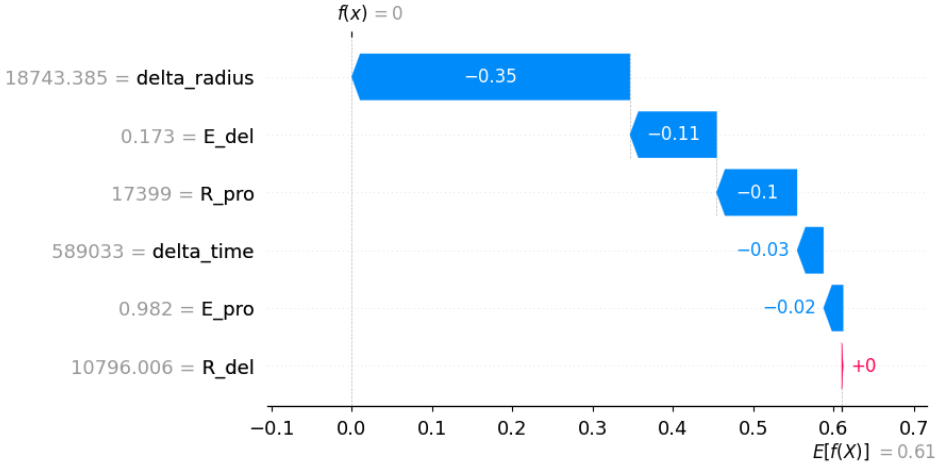
- Output XGBoost: Probabilità di classificazione per ogni evento

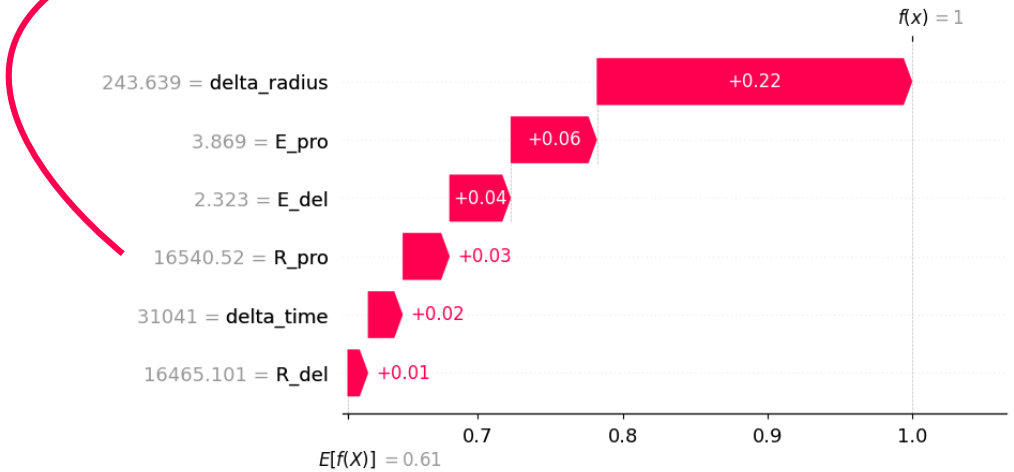


Tipico evento di BKG con probablità 0

Estrapolo due eventi

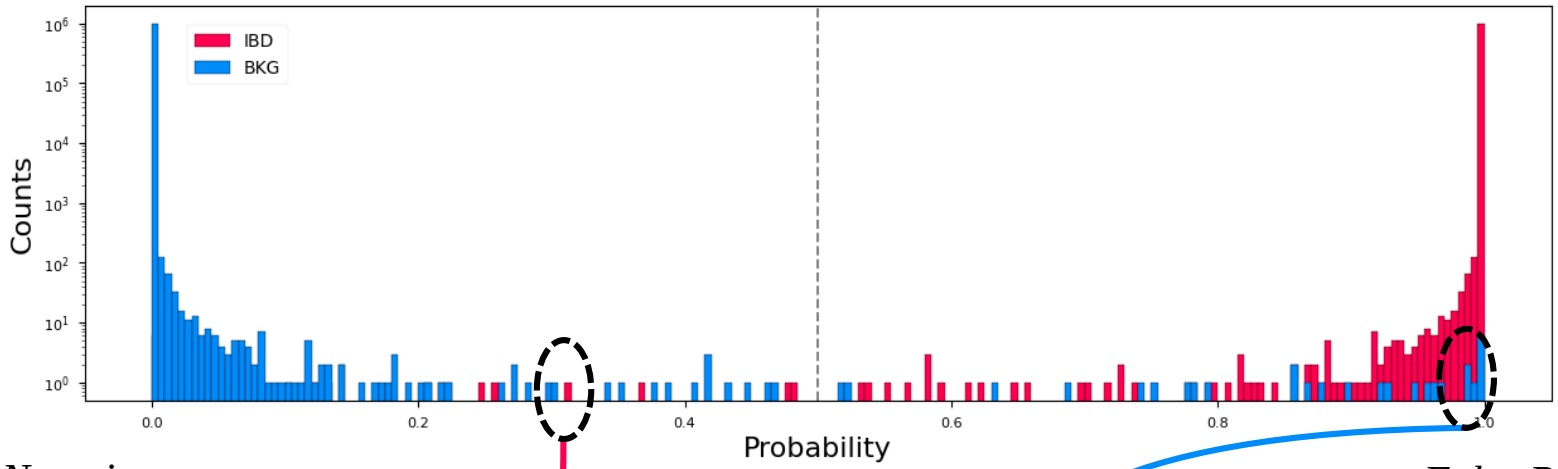
Tipico evento di IBD con probablità 1





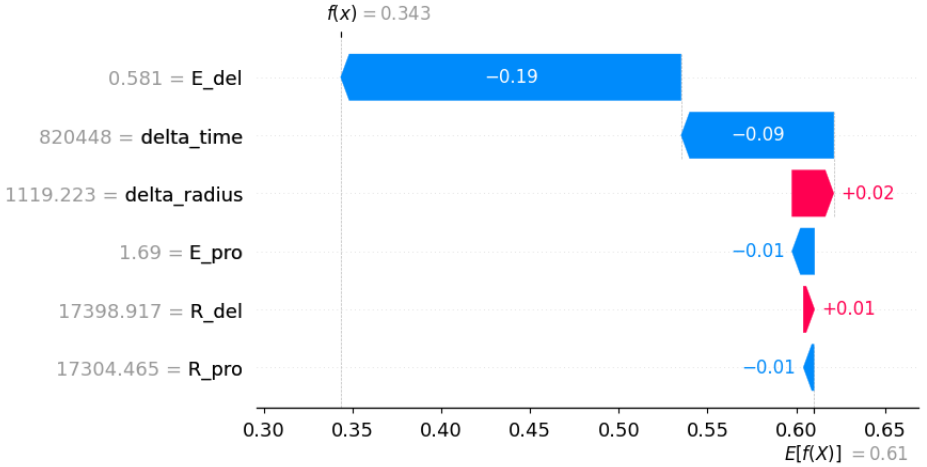
BDT: interpretazione

Estrapolo due eventi classificati errroneamente



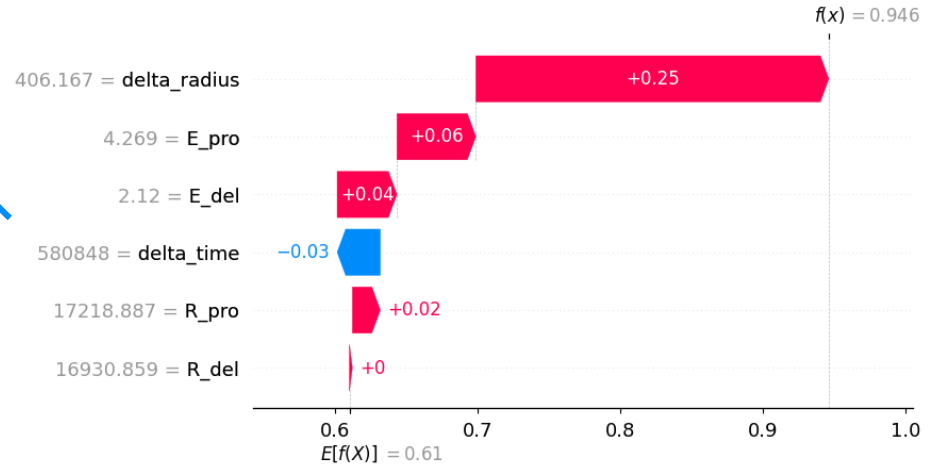
False Negative

Evento di IBD identificato come BKG



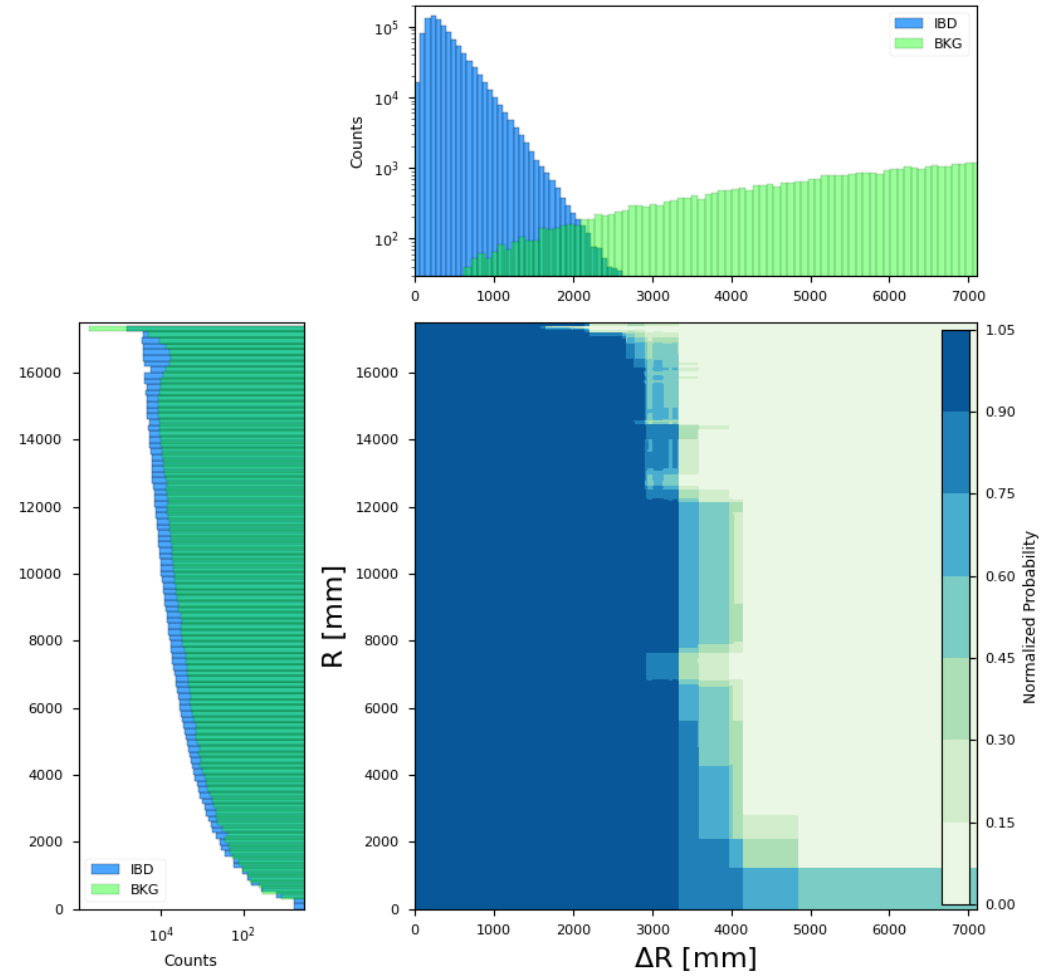
False Positive

Evento di BKG identificato come IBD



BDT: interpretazione

- Valutazione dell'algoritmo nella classifiaczione utilizzando i valori delle feature ΔR e R_{pro} .
- Approcciando il bordo del rivelatore è attesa una significativa presenza di eventi BKG attribuibili a vari materiali quali *acrilico*, *barre d'acciaio*, il *vetro dei PMT* e il *radon* naturalmente presente nella piscina d'acqua.
- L'algoritmo ha appreso l'identificazione eventi di BKG per valori di $\Delta R > 4000$ mm, in questa regione si osserva infatti una separazione tra eventi IBD e BKG.



Risultati

Risultati

- Per le valutazioni delle performance dei modelli si è utilizzato un sample formato dal 20% del dataset originario (400 033 ev)
  - XGBoost non è stato addestrato su questo sample, per evitare bias nella valutazione

Manual Cut:

	Predicted IBD	Predicted BKG
Actual IBD	194844	4542
Actual BKG	7	200640

- Mostra un numero considerevole di errori nella classificazione degli eventi IBD, con 4542 eventi IBD erroneamente identificati come BKG.
- Fornisce un efficace esclusione di eventi di BKG.

XGBoost:

	Predicted IBD	Predicted BKG
Actual IBD	199811	4
Actual BKG	3	200215

- Minor numero di errori di classificazione per eventi IBD (4 eventi FN) e BKG (3 eventi FP).
- La precisione del modello è elevata, rendendolo particolarmente affidabile per la corretta individuazione di entrambe le categorie di eventi.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Conclusioni

Conclusioni

- Investigazione di modelli per la classificazione di eventi IBD e di BKG nell'esperimento JUNO
- Migliorare il modello di riferimento attuale, l'algoritmo di classificazione 'Manual Cut', mediante l'uso di tecniche sofisticate di machine learning

Risultati: in termini di efficiency → il numero di eventi selezionati diviso il numero totale di eventi, per ogni categoria

	Manual Cut
IBD Efficiency	97.702%
BKG Efficiency	99.997%

	XGBoost
IBD Efficiency	99.9985%
BKG Efficiency	99.9979%

Bibliografia

[1] A. Abusleme et al., "Juno physics and detector", Progress in Particle and Nuclear Physics 123, 103927 (2022), https://www.sciencedirect.com/science/article/pii/S0146641021000880

[2] Origine Immagine: Li, Teng & Xia, Xin & Huang, Xingtao & Zou, JiaHeng & Li, WeiDong & Lin, Tao & Zhang, Kun & Deng, ZiYan. (2017). Design and Development of JUNO Event Data Model. Chinese Physics C. 41. 10.1088/1674-1137/41/6/066201

[3] Deng, Haowen & Zhou, Youyou & Wang, Lin & Zhang, Cheng. (2021). Ensemble learning for the early prediction of neonatal jaundice with genetic features. BMC Medical Informatics and Decision Making. 21. 10.1186/s12911-021-01701-9.

[4] A. Abusleme et al., "Sub-percent precision measurement of neutrino oscillation parameters with juno", Chinese Physics C 46, 123001 (2022), https : / / doi . org / 10 . 1088 / 1674 - 1137/ac8bc9.

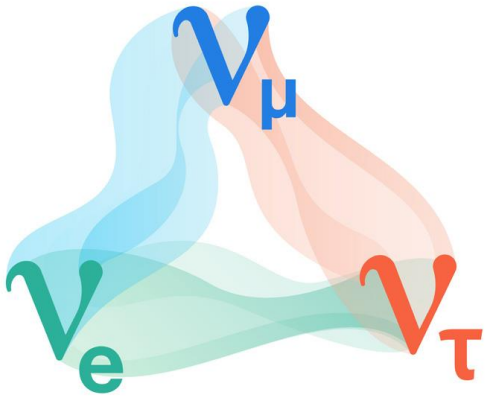
Backup

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Introduzione

- **Neutrino Oscillations**

Neutrino Oscillations



- I neutrini sono leptoni privi di carica elettrica e si presentano in tre differenti 'sapori' (v_e, v_μ, v_τ)
- Recenti esperimenti hanno dimostrato che possiedono una massa non nulla.
- *Oscillazioni dei neutrini:* un fenomeno quanto-meccanico attraverso il quale il neutrino cambia il suo "sapore" (e, μ, τ) durante la propagazione.

Ogni autostato di sapore può essere considerato come una combinazione degli autostati di massa, (ν_1,ν_2,ν_3)

La matrice Pontecorvo-Maki-Nakagawa-Sakata (PMNS) lega autostati di sapore a gli autostati di massa

Probabilità di oscillazione in funzione della distanza percorsa dal neutrino. Nel caso dell'esperimento JUNO si è interessati a $P\left(\bar{\nu}_e \rightarrow \bar{\nu}_e\right)$

$$\begin{pmatrix} v_e \\ v_\mu \\ v_\tau \end{pmatrix} = U_{\mathrm{PMNS}} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

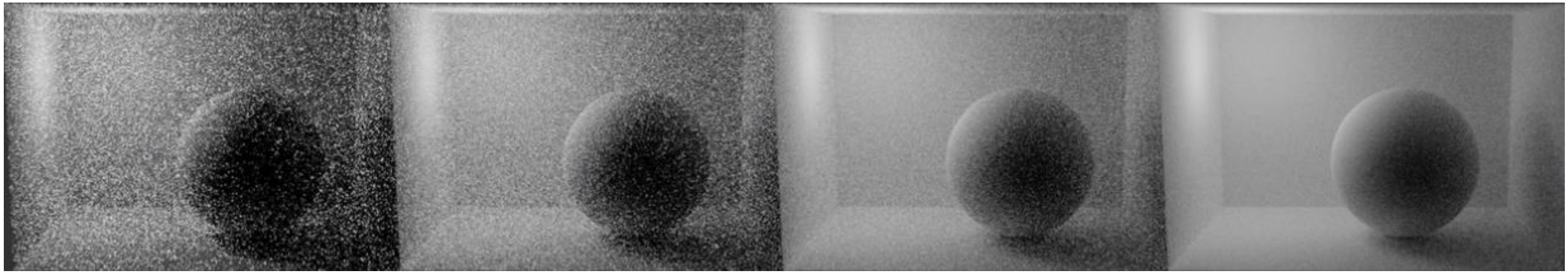
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Montecarlo

FABIO CUFINO

Metodo Montecarlo

- Il metodo Monte Carlo è una tecnica di simulazione che si basa sulla generazione casuale di numeri per risolvere problemi. Utilizzato quando il problema è troppo complesso per risolverlo con metodi analitici standard.
- La legge dei grandi numeri descrive il risultato dell'esecuzione dello stesso esperimento un gran numero di volte. Secondo la legge, la media aritmetica dei risultati ottenuti da un gran numero di prove dovrebbe avvicinarsi al valore atteso e tenderà a rimanere vicino a questo valore quanto più prove vengono condotte.
- Dato che il metodo Monte Carlo si basa su ripetute simulazioni o "prove", la legge dei grandi numeri assicura che la media dei risultati di queste simulazioni converga verso il valore atteso.



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Machine Learning

- # Supervised Learning

Supervised Learning

Il ***Machine Learning***: ramo dell'intelligenza artificiale dedicato allo sviluppo di algoritmi e modelli statistici che consentono ai computer di apprendere autonomamente dai dati, senza la necessità di una programmazione esplicita.

Supervised Learning:
L'apprendimento avviene grazie a dati etichettati

→

Dati che sono stati classificati in base a specifiche categorie o classi. L'algoritmo viene addestrato su questo set di dati e impara a riconoscere e prevedere le etichette corrispondenti a nuovi dati.

Analisi di due differenti algoritmi di Machine Learning

1. Boosted Decision Trees (BDT) ***2. Neural Networks (NN)***

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Machine Learning

- # Neural Networks

FABIO CUFINO

Neural Networks

Introduzione: Modelli computazionali ispirati al funzionamento del cervello umano.

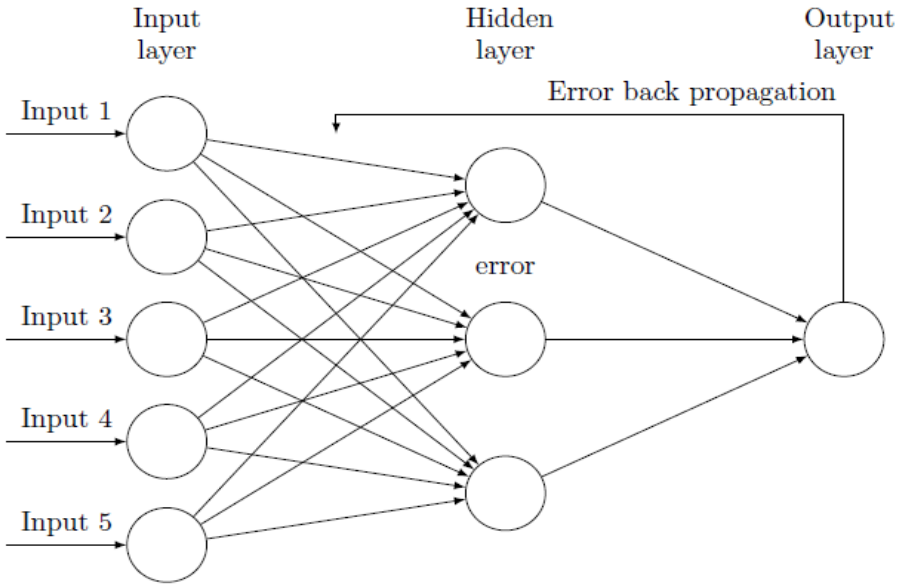
- Composte da unità di calcolo, i "neuroni", organizzate in strati.

Struttura:

- un "*input layer*",
- uno o più "*hidden layer*"
- un "*output layer*"

Ogni layer è composto da neuroni che ricevono input, eseguono calcoli e inviano l'output al prossimo strato.

Processo di Apprendimento: la rete adatta i pesi delle connessioni tra i neuroni per minimizzare l'errore tra l'output prodotto e l'output desiderato. Questo processo è chiamato "*backpropagation*"



Applicazione: Una volta addestrata, la rete neurale può essere utilizzata per prevedere l'output per nuovi input, basandosi su ciò che ha imparato dai dati di addestramento.

PyTorch: libreria per implementare la Neural Network

- **La struttura delle rete:**
- un input layer, da 6 neuroni
  - 4 layer nascosti
  - un layer di output

- Ogni hidden layer contiene 64 neuroni e utilizza la funzione di attivazione *Rectified Linear Unit (ReLU)*, definita come $ReLU(x) = \max(0, x)$

- **L'addestramento:** GPU compatibile con CUDA per sfruttare l'accelerazione hardware, utilizzando una macchina virtuale su CloudVeneto dotata di NVIDIA T4 Tensor Core GPU

- *Cross-Entropy Loss* come loss funcion

Risultati

PyTorch:

	Predicted IBD	Predicted BKG
Actual IBD	200231	46
Actual BKG	30	199726

- Incremento nel numero di eventi BKG erroneamente classificati come IBD (30 eventi),
- Elevato numero di eventi IBD erroneamente classificati come BKG (46) rispetto agli altri due modelli.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Machine Learning

- **BDT: configurazione**

BDT: configurazione

XGBoost: introduce diversi implementazioni come il parallel processing

Parametri di Configurazione dell'Algoritmo:

- **Random seed**: Configurato a 1 per assicurare la riproducibilità e garantire una generazione consistente di numeri casuali.
- **Number of estimators**: Impostato con un valore iniziale di 10.000 alberi decisionali per controllare la complessità del modello. Tuttavia, durante l'addestramento, il numero di stimatori è stato determinato automaticamente utilizzando la condizione di stop anticipato che monitora le prestazioni del modello su un set di validazione
- **Learning rate**: Configurato a 0.05, determina il contributo di ogni albero alla previsione finale.
- **Maximum tree depth**: Limitato a 3, per controllare la complessità di ciascun albero.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

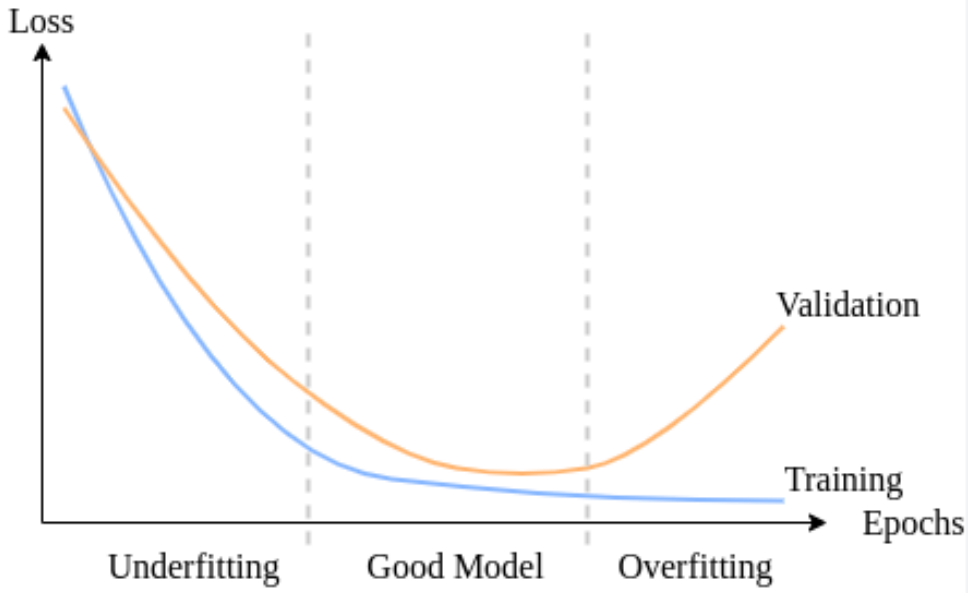
Machine Learning

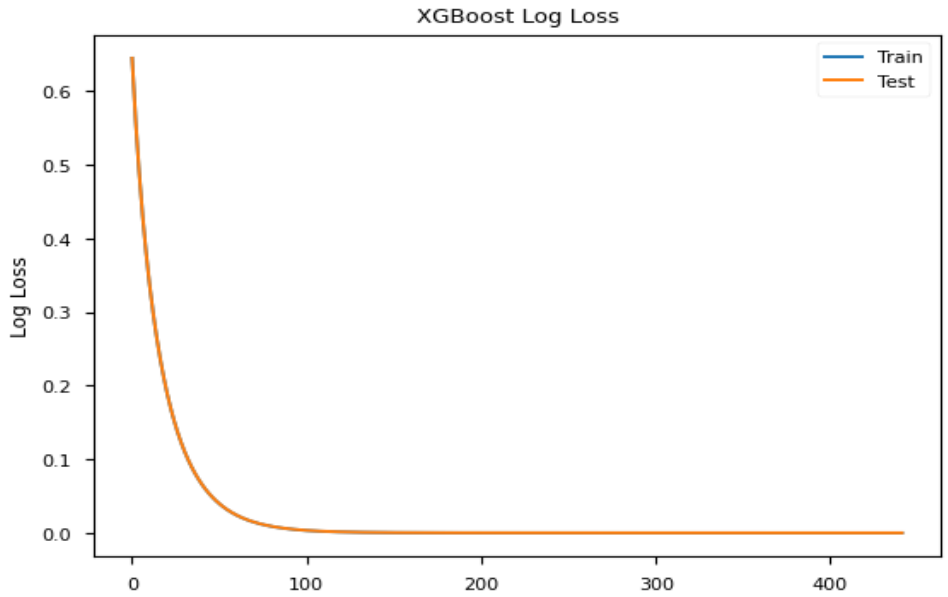
- # BDT: Overfitting e Underfitting

Università degli Studi di Padova

Overfitting si verifica quando addestriamo un modello di machine learning troppo sul set di addestramento

Underfitting si verifica quando sul modello di apprendimento automatico non è stato fatto abbastanza training sul set di addestramento





UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Background

UNIVERSITÀ DEGLI STUDI DI PADOVA

BKG cosmogenici: I raggi cosmici, soprattutto i muoni ad alta energia, interagiscono con i materiali del rivelatore, producendo neutroni veloci e isotopi instabili attraverso un processo chiamato spallazione. Gli isotopi prodotti, come Li-9, He-8 e C-11, sono instabili e decadono, contribuendo ad ulteriori eventi di fondo.

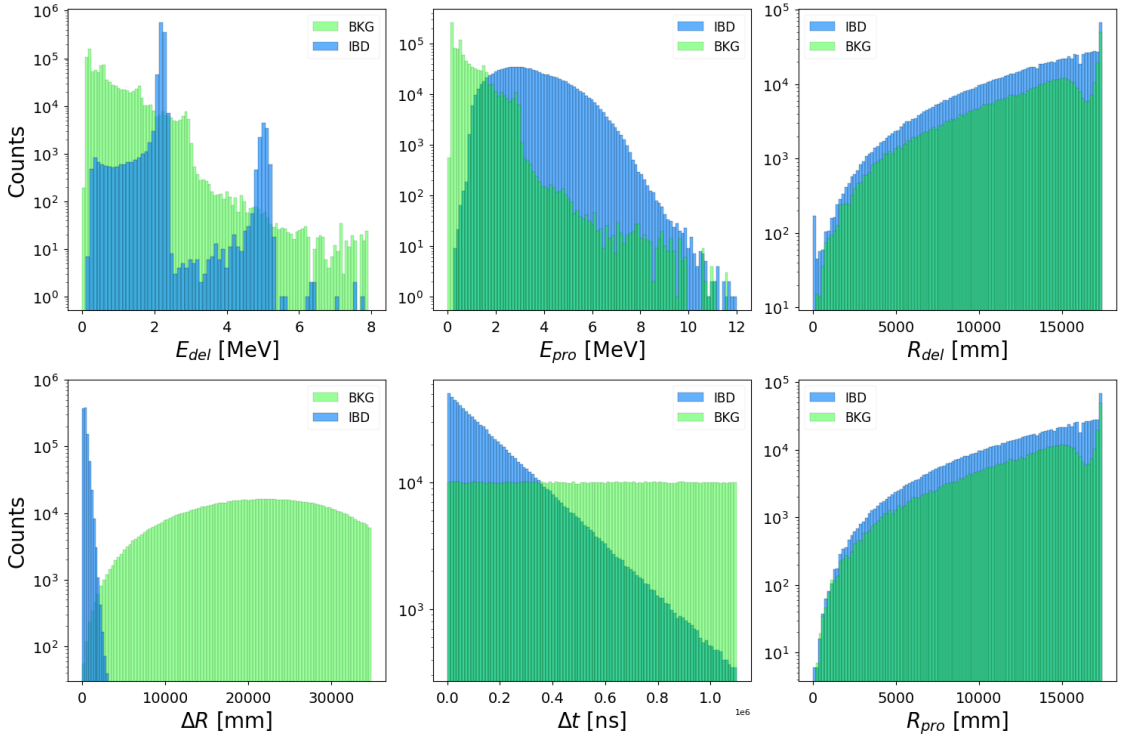
→ I neutroni veloci e gli isotopi instabili possono generare segnali che imitano un evento IBD. Questi segnali sono generati da un elettrone (che può apparire come un positrone) e da un neutrone (che può essere catturato da un protone nel rivelatore, producendo un segnale identico a quello di un neutrone in un evento IBD)

13C e particella alpha: Questo è un decadimento radiogenico che produce una particella alfa e un neutrone, imitando un evento IBD all'interno del liquido scintillante. Questo è l'unico sfondo correlato che richiede considerazione.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Feature Distribution

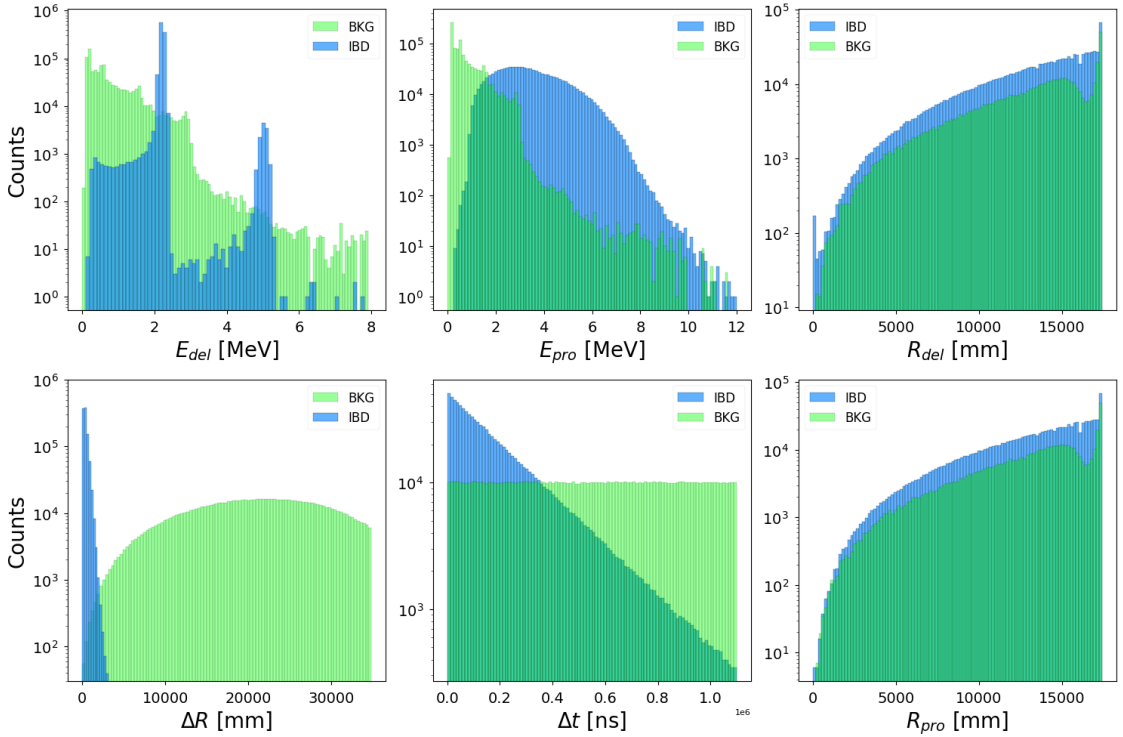
Feature Distribution

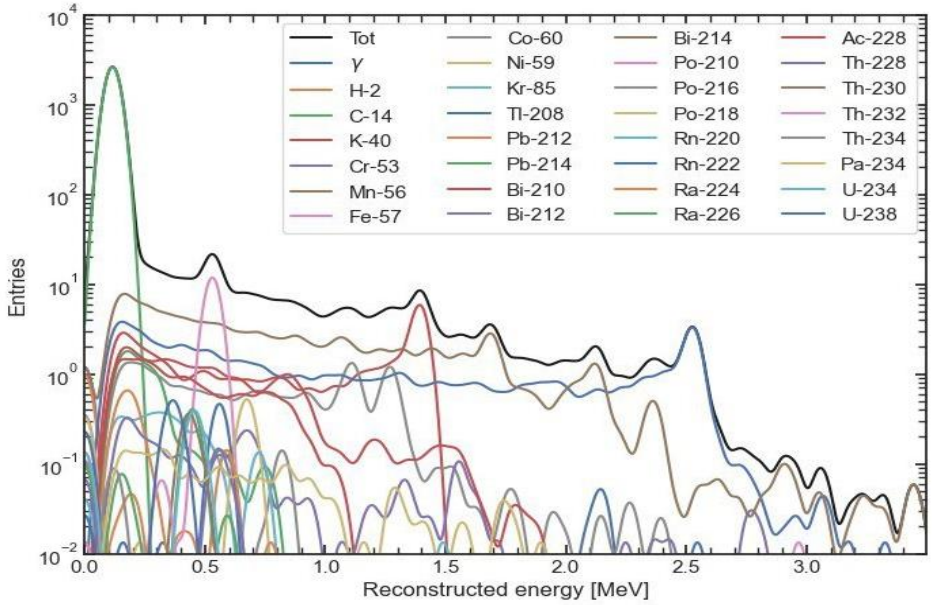


Background

Dataset Name	Number of Events	Rates (used in elecsim)
U238@LS	1,000,000 events	3.234 Hz
Th232@LS	1,000,000 events	0.733 Hz
K40@LS	1,000,000 events	0.53 Hz
Pb210@LS	1,000,000 events	17.04 Hz
C14@LS	1,000,000,000 events	3.3e4 Hz
Kr85@LS	1,000,000 events	1.163 Hz
U238@Acrylic	10,000,000 events	98.41 Hz
Th232@Acrylic	10,000,000 events	22.29 Hz
K40@Acrylic	10,000,000 events	161.25 Hz
U238@node/bar	100,000,000 events	2102.36 Hz
Th232@node/bar	100,000,000 events	1428.57 Hz
K40@node/bar	100,000,000 events	344.5 Hz
Co60@node/bar	100,000,000 events	97.5 Hz
U238@PMTGlass	1,000,000,000 events	4.90e6 Hz
Th232@PMTGlass	1,000,000,000 events	8.64e5 Hz
K40@PMTGlass	1,000,000,000 events	4.44e5 Hz
Tl208@PMTGlass	1,000,000,000 events	1.39e5 Hz
Co60@Truss	0	? Hz
Tl208@Truss	0	? Hz
Rn222@WaterRadon	100,000,000 events	90 Hz

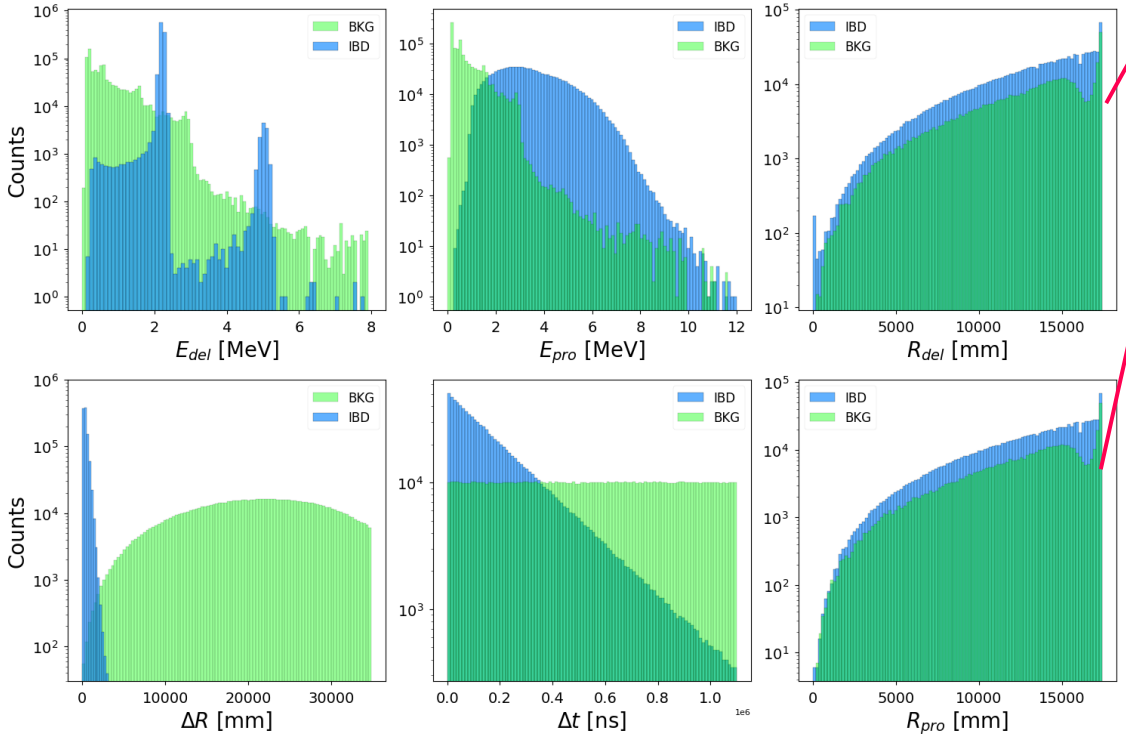
Feature Distribution





Feature Distribution

Il carbonio-14 subisce un decadimento beta- e decade quindi per emissione di un elettrone e antineutrino dando origine al nucleo figlio di azoto-14



Gran parte degli eventi presenti al centro del detector e quindi uniformemente distribuiti (e non dovuti alla radioattività esterna di acrilico e PMT) è data dal C14.

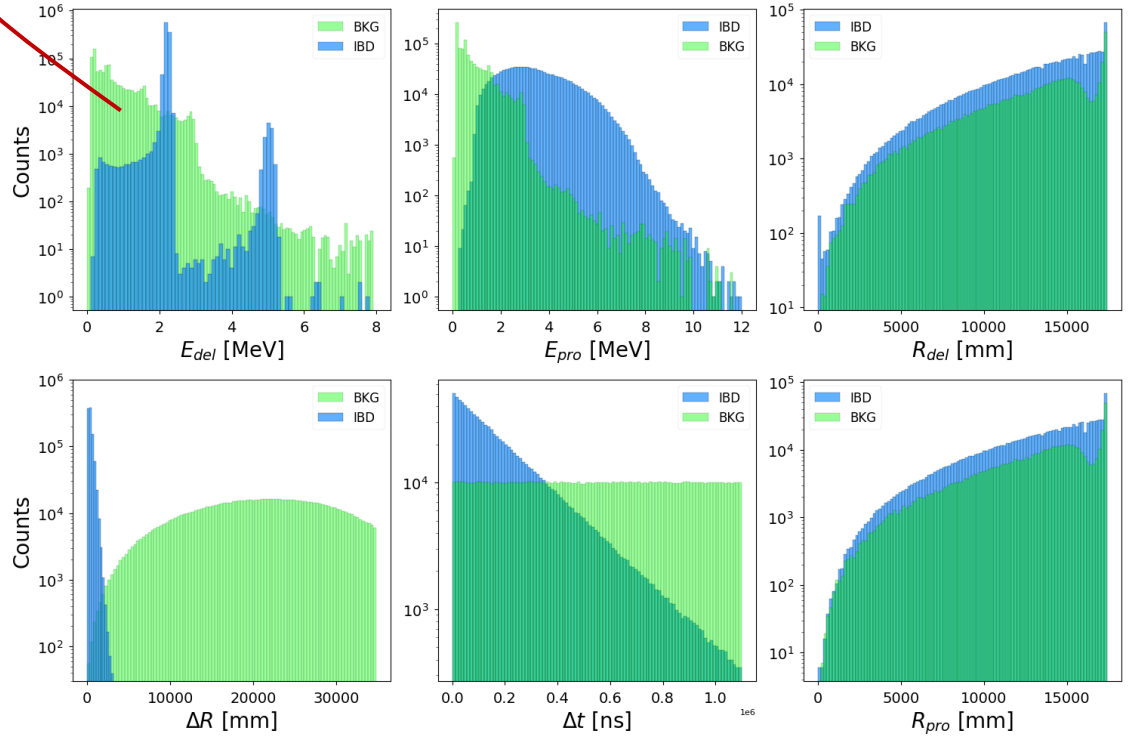
Il decadimento del C14 ha un'energia molto appena sufficiente ad attivare il numero minimo di PMT necessari al trigger.

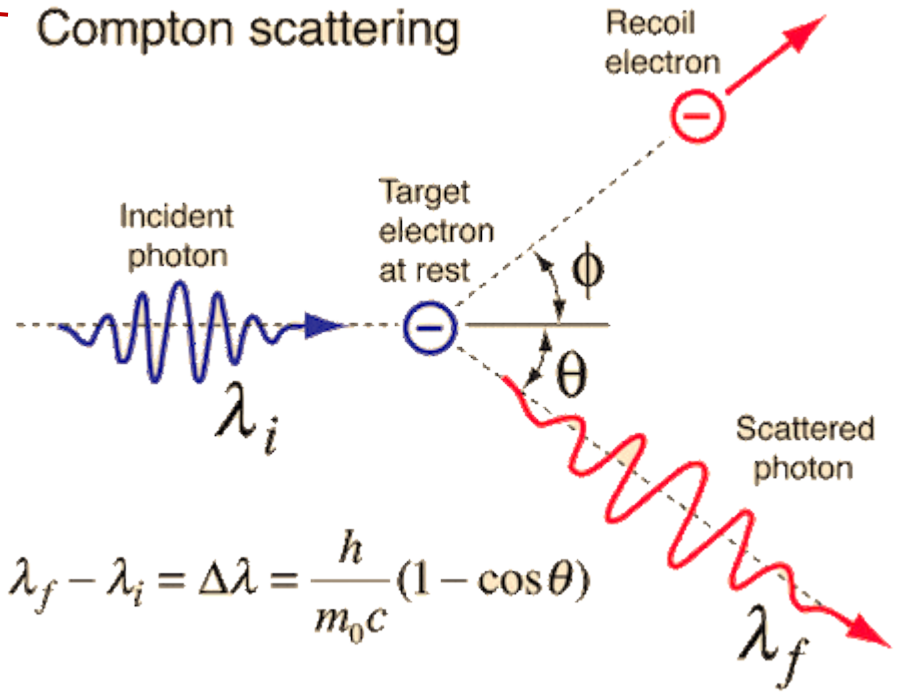
Spostandosi verso il bordo del rivelatore gli eventi di C14 attivano sempre meno PMT, perché la luce deve arrivare fino alla parte opposta del detector.

Di conseguenza solo alcuni di questi eventi accendono abbastanza PMT e vengono rilevati.

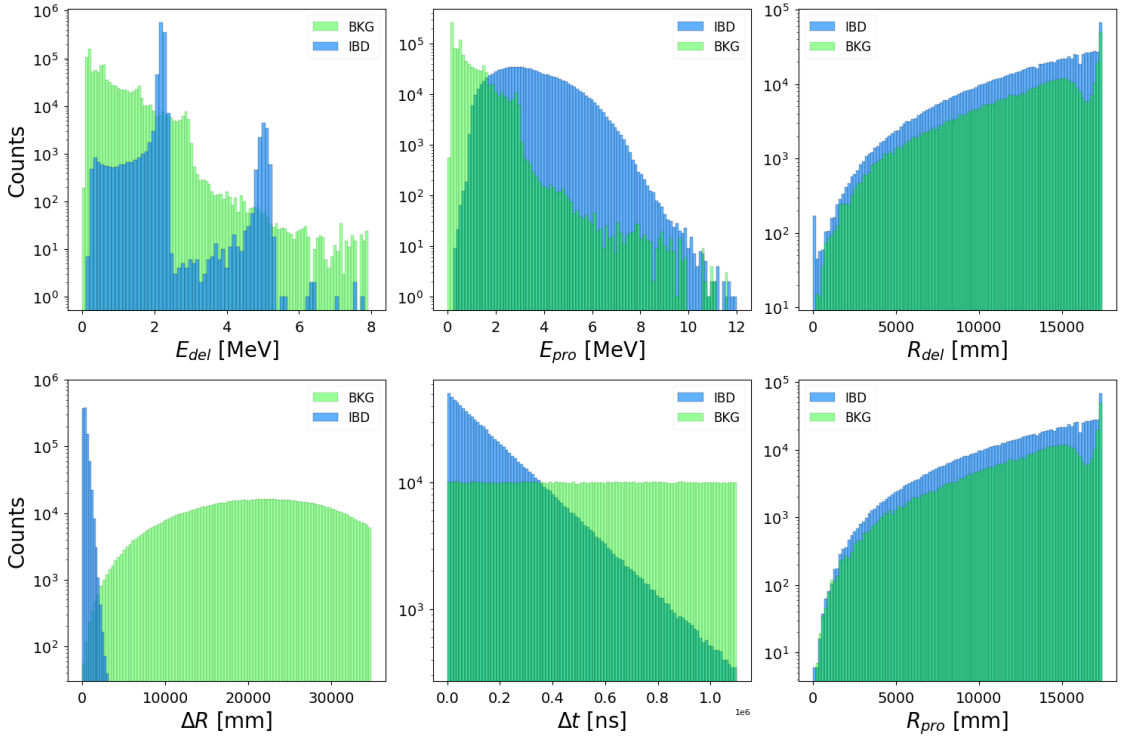
Risultato: Meno eventi vicino al bordo

Feature Distribution





Feature Distribution



$$N(t) = N_0 \, e^{-\lambda t}$$

$$\tau = \frac{1}{\lambda}$$

τ = vita media

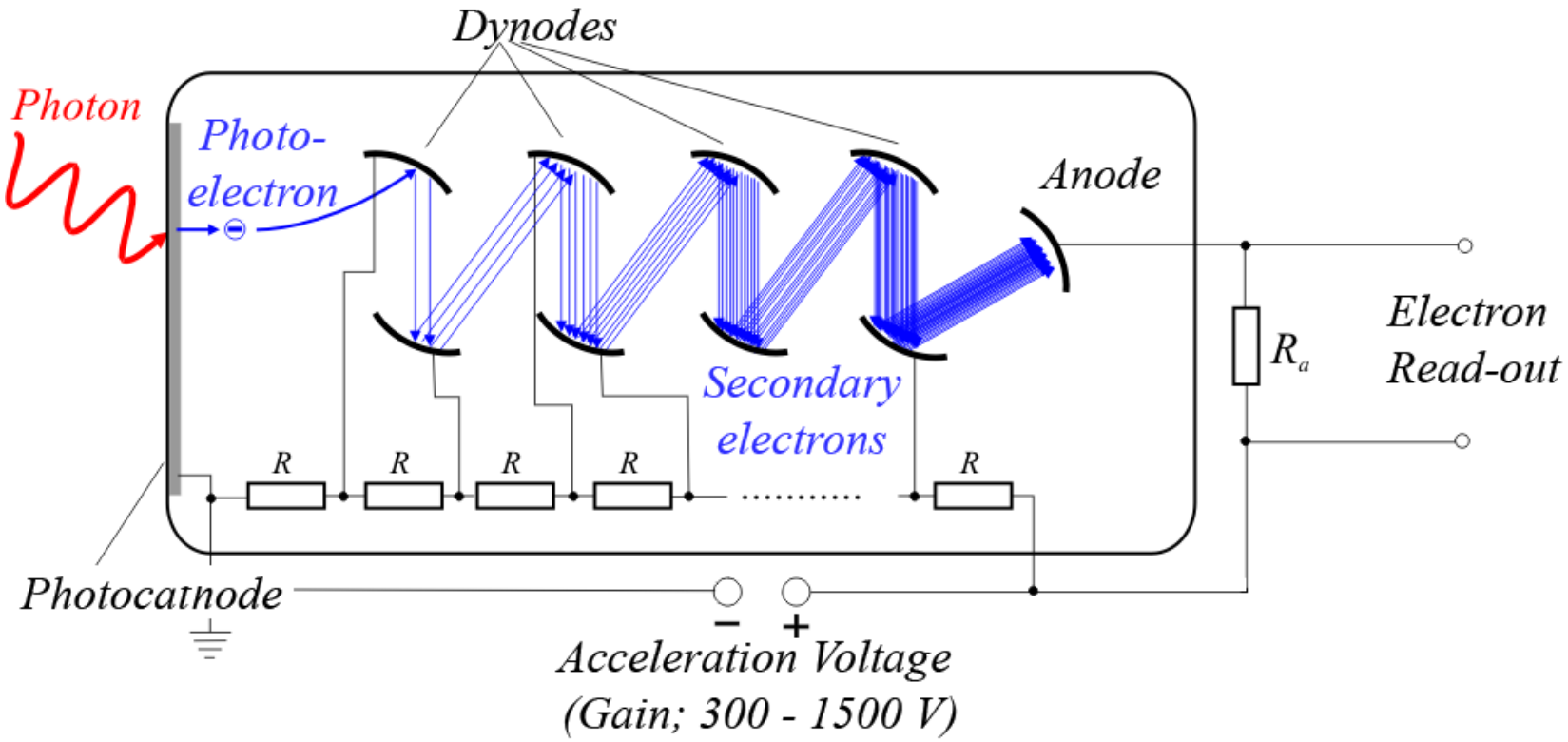
λ = costante di decadimento

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

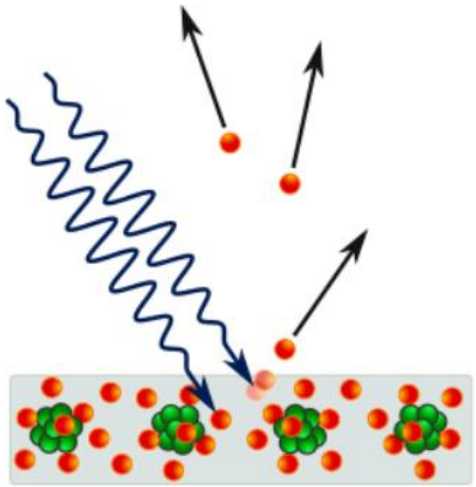
PMT

FABIO CUFINO

Università degli Studi di Padova



Effetto fotoelettrico:
una lastra metallica conduttrice investita da una radiazione UV si carica positivamente

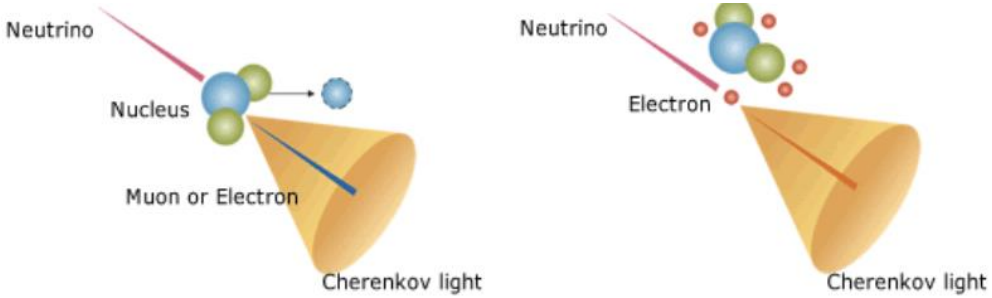


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

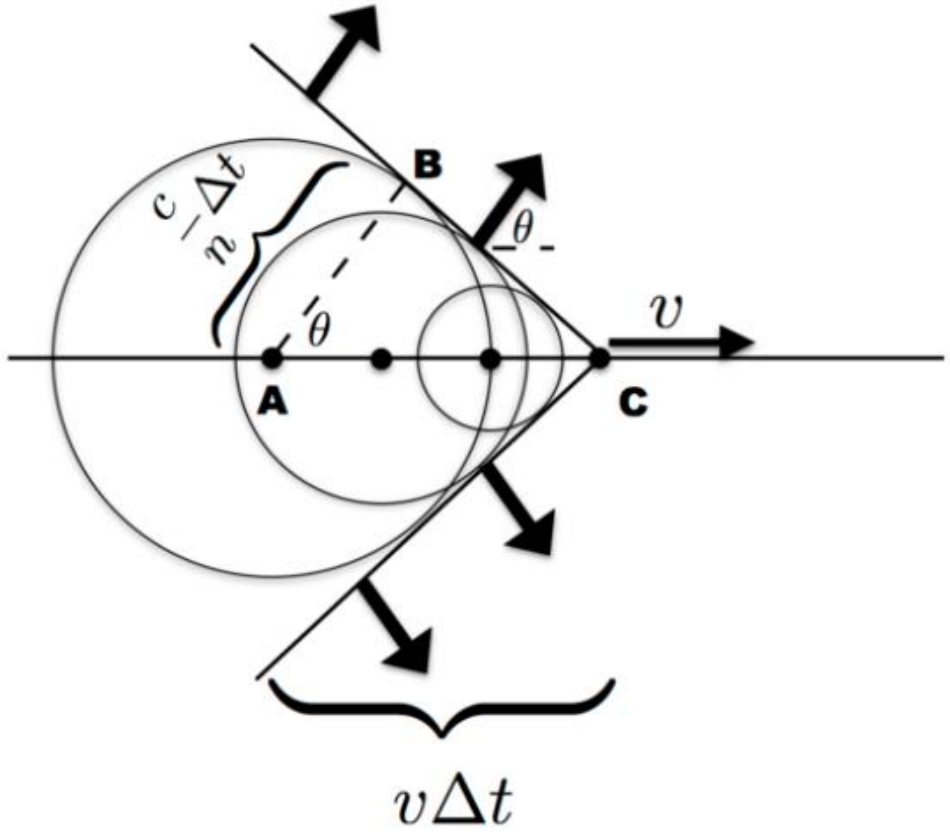
Rivelatore Cherenkov

FABIO CUFINO

Rivelatore Cherenkov



The generated charged particle emits the Cherenkov light.

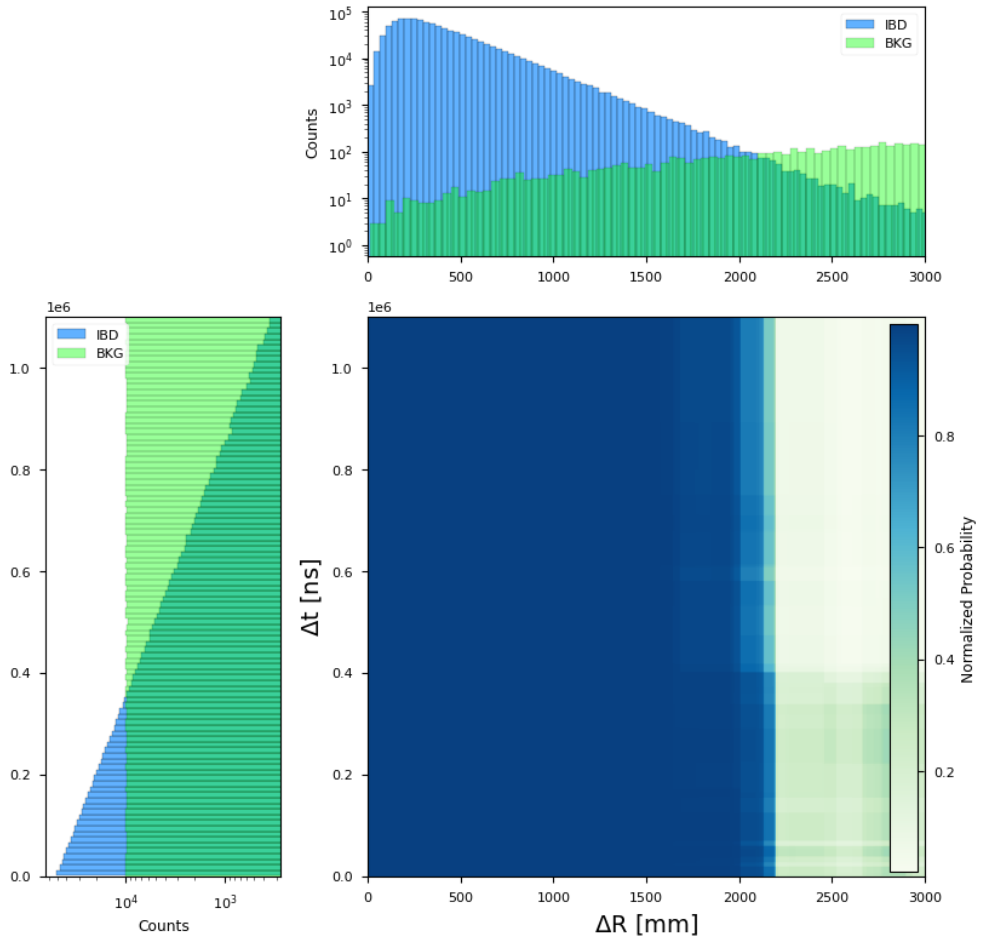


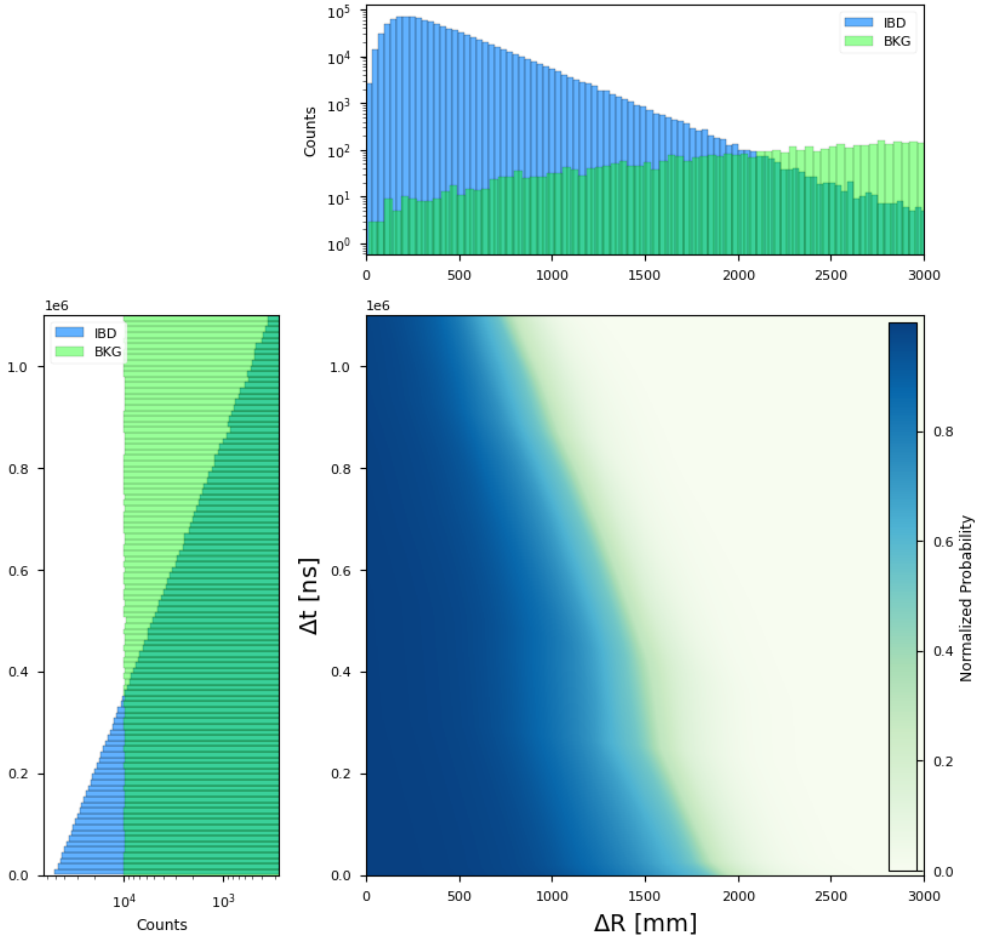
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Model Comparison

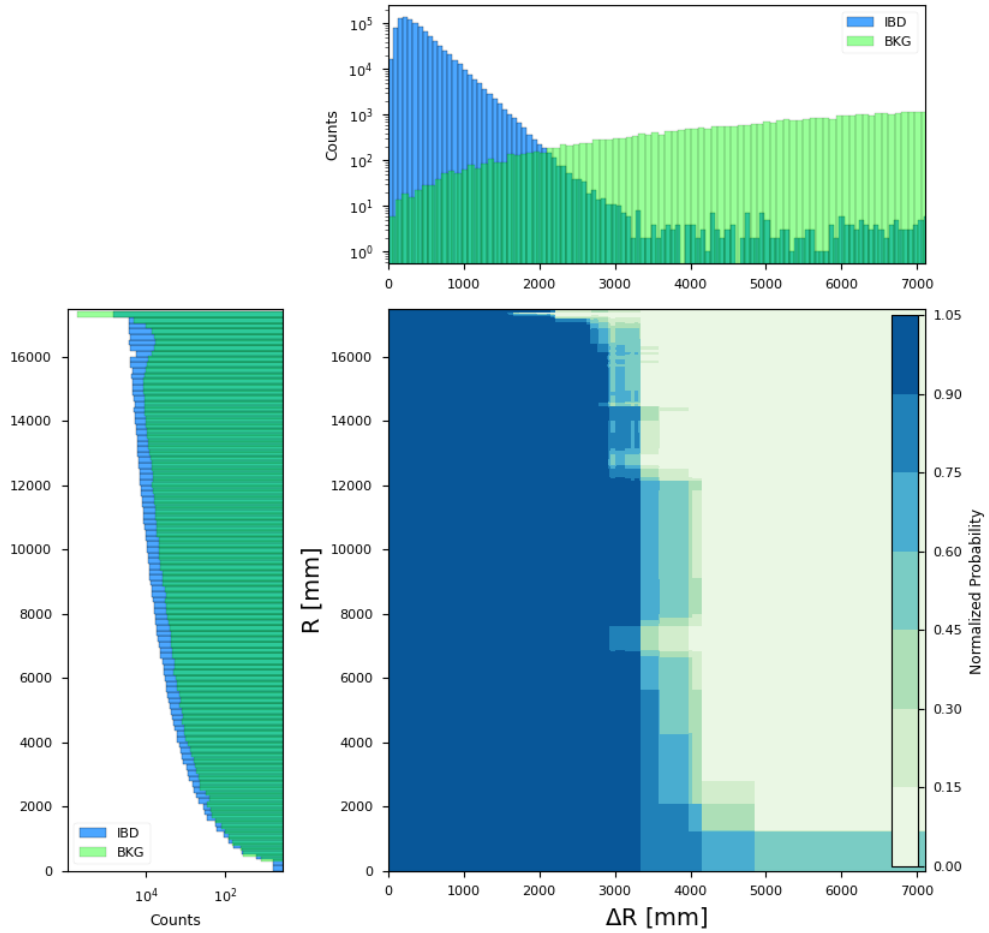
- **BDT v.s. Neural Network**

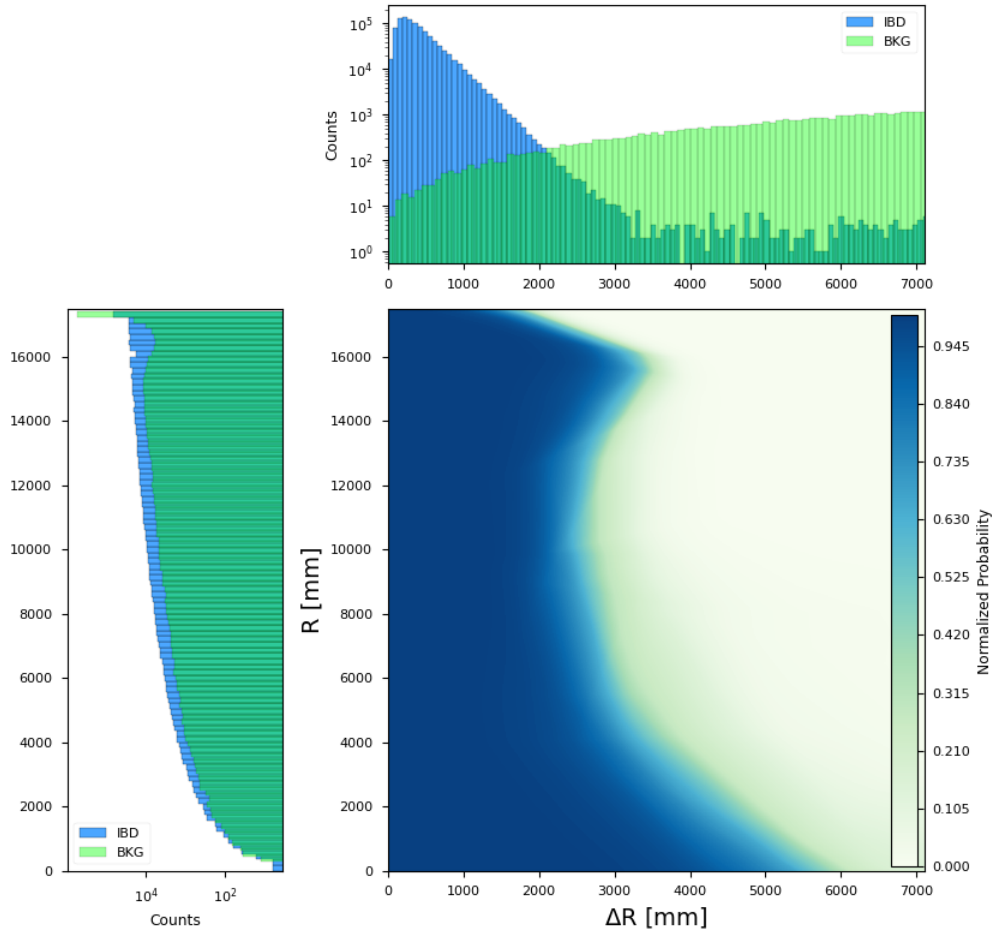
Model Comparison





Model Comparison





Efficiency and Purity

Valutati *l'efficiency* e la *purity* della selezione.

Calcolate in base al numero di eventi al giorno di, "Correlated Background " e "Accidental Background".

IBD attesi: Per ogni sorgente, calcolato come l'IBD efficiency moltiplicata per il muon cut e per ev/day.

IBD attesi da Accidentals: Ottenuti moltiplicando il muon cut, l'ev/day e la 1 - BKG efficiency

	Manual Cut	XGBoost	Neural Network
Purity	0.6610	0.7054	0.1981
Efficiency	0.8949	0.9160	0.9158

	ev/day	muon cut	Manual Cut	XGBoost	PyTorch
Reactor	57.4	0.916	51.4	52.6	52.6
Geo-U	1.155	0.916	1.03	1.06	1.06
Geo-Th	0.345	0.916	0.31	0.32	0.32
Li9	0.81	1	0.79	0.81	0.81
He8	0.09	1	0.09	0.09	0.09
World Reactors	1.22	0.916	1.09	1.12	1.12
Atmospheric ν	0.2	0.916	0.18	0.18	0.18
Fast neutron	0.12	0.916	0.11	0.11	0.11
$C(\alpha, n)^{16}O$	0.06	0.916	0.05	0.05	0.05
Total	–	–	55.02	56.32	56.31

	ev/day	muon cut	Manual Cut	XGBoost	PyTorch
Accidentals	134124.0	0.916	22.69	18.22	209.01

Efficiency: Calcolata come il numero di IBD selezionati diviso per il totale di IBD.

Purity: Calcolata come il numero di IBD selezionati diviso per il totale degli eventi selezionati.