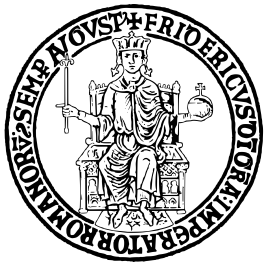
UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"



Scuola Politecnica e delle Scienze di Base

Area Didattica di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica "Ettore Pancini"

Laurea Triennale in Fisica

Ricostruzione di bosoni di Higgs in coppie di fotoni con algoritmi di Machine Learning per ricerche di nuova fisica col rivelatore CMS ad LHC

Relatori:
Prof. Alberto Orso Maria Iorio

Candidato:
Gaia Mattia
Matr. $N85000888$

Anno Accademico 2019/2020

Indice

Introduzione

Il Modello Standard (MS) è la teoria fisica che ad oggi descrive meglio le componenti elementari della materia e tre delle quattro interazioni fondamentali della natura. Esso lascia ancora alcuni interrogativi aperti: oltre a non trovare una collocazione all'interazione gravitazionale, non spiega altri fenomeni come le incongruenze tra teoria ed esperimenti su cui si fonda il *problema della gerarchia*, non chiarisce come mai ci sono tre famiglie di particelle per i leptoni e i quark né garantisce che queste particelle siano elementari.
Tutti questi interrogativi aperti, assieme a numerosi altri, hanno portato a sviluppare numerosi modelli, a cui ci si riferisce quando si parla di teorie *oltre il modello standard* (BSM, acronimo di *Beyond Standard Model)*. In alcune di queste è ipotizzata l'esistenza di particelle chiamate *Vector Like Quark* (VLQ): fermioni la cui massa si suppone essere dell'ordine dei TeV.
Il laboratorio più adatto per l'indagine sperimentale di nuova fisica, dunque anche della rivelazione dei VLQ, è il *Large Hadron Collider* (LHC), un sincrotrone situato presso il CERN, con 13 TeV di energia disponibili per collisione protone-protone.
Compact Muon Solenoid (CMS) è un sistema di rivelatori situato in uno dei punti di collisione di particelle di LHC che ha lo scopo di ricercare i limiti del MS e i possibili modelli di nuova fisica. Nel 2012, insieme all'esperimento ATLAS, esso ha confermato l'esistenza del bosone di Higgs, già teorizzata decenni prima nell'ambito del MS.
L'elaborazione dei dati di CMS è un l'ultimo passaggio cruciale per validare o smentire le previsioni dei modelli teorici, siano essi SM o BSM. Anche dopo la drastica riduzione dei dati effettuata dall'hardware degli esperimenti ad LHC, la banca dati del CERN elabora in media un petabyte di dati al giorno; in più, molti dei processi analizzati sono complessi da ricostruire, spesso anche a causa della compresenza di molteplici fondi che competono alla medesima segnatura del detector. Per l'analisi di tali dati è sempre più importante e diffuso in fisica delle alte energie l'utilizzo di tecniche di *Machine Learning*, in modo da rendere più veloci ed efficienti le procedure di ricostruzione delle particelle.
Il presente elaborato si pone come obiettivo la ricostruzione del bosone di Higgs

derivante dal canale di decadimento $H \to \gamma\gamma$ mediante l'utilizzo di un algoritmo supervisionato. Per tale scopo, sono stati utilizzati dei dati derivanti da simulazioni di collisioni p-p in CMS alle condizioni di presa dati della Run II di LHC in ambito MS e BSM. Per quest'ultimo è stata considerata una produzione singola dell'ipotetico *vector like quark* T, facendo varie ipotesi sui valori della sua massa, in particolare studiando il suo canale di decadimento $T \to Ht$. Per i campioni SM il processo di produzione dell'Higgs è accompagnato da un quark top e un quark leggero .
Questa tesi è organizzata come segue:

- Capitolo 1: viene fatta un'introduzione al Modello standard e un accenno ai modelli proposti come soluzione dei suoi limiti;
- Capitolo 2: descrizione di LHC e in particolare dell'esperimento CMS;
- Capitolo 3: dopo una breve introduzione al *Machine Learning*, è delineato il processo utilizzato per la ricostruzione dei bosoni di Higgs dai campioni presi in esame così da generare l'input per allenare un algoritmo supervisionato; infine vengono riportati i risultati dell'allenamento.

Capitolo 1

Modello Standard e teorie oltre il Modello Standard

1.1 Le particelle elementari

Una descrizione qualitativa della materia ordinaria la vede costituita in buona parte dai nuclei degli atomi, separati da una distanza dell'ordine dell'angstrom ($10^{-10}\ m$), molto maggiore rispetto al loro raggio, che invece è dell'ordine del fermi ($10^{-15}\ m$). I nuclei atomici non sono particelle elementari, bensì sono composti da stati legati di particelle elementari.
La natura delle particelle elementari e la forma dell'interazione tra di esse sono oggetti di studio della fisica delle particelle. Effettuando un paragone qualitativo con la fisica macroscopica, per studiarle si potrebbero porre le particelle a distanze relative differenti e misurare la forza tra loro, analogamente a come ha fatto Coulomb per la forza elettrica o Cavendish per la gravità. Tuttavia, essendo il mondo microscopico governato da leggi totalmente diverse rispetto a quello macroscopico, la formulazione di un modello che lo descriva al meglio non è più guidata dalla fisica classica.
Proprio come la luce, anche la materia ha proprietà ondulatorie: questa proprietà della natura è descritta dalla meccanica quantistica, che sostituisce la meccanica classica in scala di lunghezza o di energia atomica e subatomica. Inoltre, per gli oggetti che viaggiano a velocità paragonabili a quella della luce, le leggi che regolano lo spazio tempo e le composizioni delle velocità sono modificate dalla relatività speciale, la quale prevede la conversione di energia in materia e viceversa. Al livello microscopico, ciò rende possibile processi di creazione e distruzione di particelle di materia.
In sintesi, per descrivere le particelle elementari, è necessaria una teoria che incorpori sia la relatività che i principi quantistici: la teoria quantistica dei campi.

Velocità vicina a $c \downarrow$ \ Scala più piccola $\longrightarrow$	Meccanica classica	Meccanica quantistica
	Meccanica relativistica	Teoria quantistica dei campi

I dettagli di come tale formalismo si applica alla realtà possono solo essere dati dagli esperimenti. Ad oggi, la teoria quantistica dei campi che meglio descrive il mondo delle particelle è il cosiddetto Modello Standard, abbreviato in SM (*Standard Model*), la cui prima formulazione risale agli anni cinquanta, per poi consolidarsi nel corso dei decenni successivi.
Le interazioni fondamentali finora note sono quattro e sono, in ordine di intensità: nucleare forte, elettromagnetica, nucleare debole e gravitazionale. Quest'ultima, che è immanente nel mondo macroscopico, non ha trovato il suo spazio nello SM, considerata trascurabile in scala microscopica a cui in genere sono studiate le altre tre. Questo è uno, ma non l'unico, degli interrogativi aperti che rendono il Modello Standard una teoria incompleta, in quanto non capace di spiegare tutti i fenomeni fondamentali della natura.

1.2 Modello standard

Alla base della formulazione del Modello standard viene posto un principio di simmetria che consiste nell'invarianza della teoria sotto opportune trasformazioni locali, dette trasformazioni di gauge. Le interazioni fondamentali vengono rappresentate nel gruppo unitario $SU(2) \times U(1) \times SU(3)$, costituito dal prodotto di $SU(2) \times U(1)$, che descrive le *interazioni elettromagnetiche* e *deboli* (unificate nell'interazione *elettrodebole*), e da $SU(3)$, che descrive le interazioni forti. La descrizione delle interazioni elettromagnetiche attraverso il gruppo $U(1)$ prende il nome di **elettrodinamica quantistica**, mentre la descrizione delle interazioni forti attraverso il gruppo $SU(3)$ prende il nome di **cromodinamica quantistica**. A ogni gruppo considerato corrispondono dei bosoni vettori, che sono i mediatori delle interazioni [1].
Il Modello Standard dunque unifica interazione forte, debole ed elettromagnetica, fornendo un *set* di informazioni precise su ogni particella in base a caratteristiche quali numeri quantici, carica, massa, etc. Gli strumenti principali utilizzati per descrivere la probabilità di avere una certa interazione fra le particelle sono gli integrali di Feynman, che possono essere espressi anche sotto forma di diagrammi che rappresentano i termini della serie perturbativa dell'ampiezza di scattering per un processo con definiti stati iniziali e finali. La somma di tutti gli infiniti ordini perturbativi rappresenta il processo fisico effettivo.
Il numero quantico che fornisce la prima suddivisione di particelle è lo spin, che per il teorema spin-statistica, è legato al tipo di statistica a cui obbedisce la par-

ticella considerata: tutte le particelle a spin semintero sono definite **fermioni**, in quanto obbediscono alla statistica di Fermi-Dirac, quelle con spin intero sono invece chiamate **bosoni** in quanto obbediscono alla statistica di Bose-Einstein, i quali non sono vincolati al principio di esclusione di Pauli. I fermioni sono a loro volta suddivisi in **quark** e **leptoni**, che sono i costituenti della materia "ordinaria" di cui abbiamo esperienza quotidiana.
Per ogni particella esiste la corrispettiva antiparticella, che ha gli stessi numeri numeri quantici ma di segno opposto.
Esistono sei leptoni (e corrispettivi anti-leptoni) raggruppati in tre famiglie (o generazioni) differenti: elettrone e neutrino elettronico, muone e neutrino muonico e infine tauone e neutrino tauonico; ognuno contraddistinto in base alla carica Q, la massa e il numero quantico leptonico di famiglia (elettronico L_e, muonico L_μ e tauonico L_τ), schematizzati nella tabella 1.1.

Leptone	Massa MeV/c^2	Q/e	L_e	L_μ	L_τ
e^-	$\backsim 0,511$	-1	1	0	0
ν_e	$< 2 \cdot 10^{-6}$	0	1	0	0
μ^-	$\backsim 105,67$	-1	0	1	0
ν_μ	$< 1,7$	0	0	1	0
τ^-	$\backsim 1.776,8$	-1	0	0	1
ν_τ	$< 15,5$	0	0	0	1

Tabella 1.1: Classificazione dei leptoni per famiglia, corrispettiva carica, massa e numeri leptonici.

In modo del tutto analogo, esistono sei quark (e corrispettivi anti-quark) contraddistinti dalla carica, la massa e il sapore, tabulati in 1.2. Si possono suddividere in tre famiglie: up e down, charm e strange e top e bottom. Per gli ultimi quattro si definisce un numero quantico di sapore, rispettivamente: *charm* (C), *strangeness* (S), *truth* (T) e *beauty* (B). Per coerenza dovrebbero esserci anche *upness* (U) e *downness* (D), ma sarebbero ridondanti, in quanto l'unico quark con $S = C = B = T = 0$ e $Q = \frac{2}{3}$ è il quark up, quindi non è necessario specificare anche $U = 1$ e $D = 0$.
I quark sono anche contraddistinti da un altro numero quantico, la cosiddetta *carica di colore*, che può essere: rosso, blu e verde per i quark e anti-rosso, anti-blu e anti-verde per gli anti-quark. Oltretutto tali particelle, eccetto il quark top, non sono mai state osservate isolate, ma solo all'interno di particelle composite, chiamate **adroni**, a loro volta suddivisi in **barioni**, che hanno un numero dispari di quark costituenti, e **mesoni**, che ne hanno un numero pari. Per i barioni si definisce un numero quantico, chiamato numero barionico B, pari a 1 per i barioni, -1 per gli antibarioni e 0 per tutte le altre particelle.

Quark	Massa MeV/c^2	Q/e	S	C	B	T
d	$\backsim 4,8$	$-\frac{1}{3}$	0	0	0	0
u	$\backsim 2,4$	$\frac{2}{3}$	0	0	0	0
s	$\backsim 95$	$-\frac{1}{3}$	-1	0	0	0
c	$\backsim 1,27 \cdot 10^3$	$\frac{2}{3}$	0	1	0	0
b	$\backsim 4,18 \cdot 10^3$	$-\frac{1}{3}$	0	0	-1	0
t	$\backsim 172,44 \cdot 10^3$	$\frac{2}{3}$	0	0	0	1

Tabella 1.2: Classificazione dei quark per famiglia, corrispettiva carica, massa e numeri quantici di sapore.

Come accennato, i *bosoni vettori* (tabella 1.3) sono i mediatori delle interazioni: il *fotone* per la forza elettromagnetica, due *bosoni W*, con carica positiva o negativa, e un *bosone Z*, con carica neutra, per la forza debole e infine il *gluone* per la forza forte. A differenza dell'interazione elettromagnetica che è mediata dal fotone, il quale è elettricamente neutro, i gluoni trasportano colore. Essi si rivelano sperimentalmente solo all'interno di particelle composite o in combinazioni incolori con altri gluoni (*glueballs*).

Bosone	Simbolo	Interazione	Massa $(\frac{GeV}{c^2})$
gluone	g	forte	0
fotone	γ	elettromagnetica	0
bosone W	W^+, W^-	debole carica	$\backsim 80,39$
bosone Z	Z	debole neutra	$\backsim 91,19$

Tabella 1.3: Classificazione dei bosoni vettori (o bosoni di gauge).

In ultimo, ma non per importanza, c'è il *bosone di Higgs*, bosone scalare che svolge un fondamentale ruolo nell'ambito dello SM, poiché attraverso il campo di Higgs ad esso associato avviene ciò che viene chiamata rottura spontanea della simmetria di gauge elettrodebole, fenomeno che permette di conferire massa ai bosoni W e Z e ai fermioni. Tale particella, inclusa nel Modello standard dal 1967, è stata osservata per la prima volta nel 2012 grazie agli esperimenti ATLAS e CMS presso l'acceleratore LHC del CERN.

1.2.1 Elettrodinamica quantistica

L'elettrodinamica quantistica, abbreviata QED dall'inglese *Quantum-Electrodynamics*, è la meno recente delle teorie dinamiche utilizzate nel Modello Standard. L'elettromagnetismo è infatti caratterizzato dall'invarianza di tutte le grandezze osservabili e si ha un tipo di invarianza di gauge particolarmente semplice (abeliana),

mentre nello SM in generale ci saranno simmetrie di gauge non abeliane.
Tutti i fenomeni elettromagnetici sono riconducibili ad un vertice primitivo del tipo in figura 1.1.

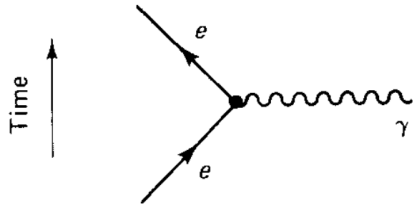


Figura 1.1: Singolo vertice di un diagramma di Feynman che rappresenta l'interazione elettromagnetica tra due *e*.

Interazioni più complesse sono rappresentabili tipicamente combinando due o più repliche di vertici primitivi e coinvolgendo particelle diverse in modo da definire gli stati iniziali e finali del processo. L'interazione tra due elettroni, ad esempio, che in elettromagnetismo classico interpretiamo come repulsione coulombiana, si può rappresentare come in figura 1.2.

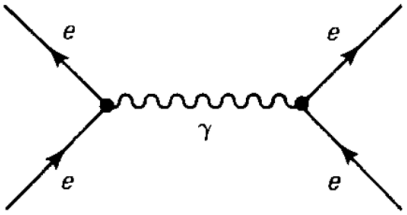


Figura 1.2: Diagramma di Feynman del *Moller scattering*.

In QED questo processo è chiamato *Moller scattering* e si interpreta l'interazione come *mediata dallo scambio di un fotone*. Questo è un esempio di diagramma di Feynman in QED. Ogni vertice all'interno di un diagramma introduce un fattore di $\alpha = \frac{e^2}{c\hbar} = \frac{1}{137}$, la costante di struttura fine, che, essendo un numero puro in valore assoluto minore di 1, comporta che i diagrammi con più vertici contribuiscono sempre meno man mano che l'ordine perturbativo aumenta e, a seconda della misura con cui si confronta, tale ordine può essere più o meno rilevante.

1.2.2 Cromodinamica quantistica

La parte dello SM che descrive l'interazione forte, cioè l'interazione tra quark mediante i gluoni, è la cromodinamica quantistica, anche abbreviata QCD, dall'inglese *Quantum-ChromoDynamics*. Poco dopo la proposta dell'esistenza dei

quark, fu introdotto il concetto di carica di colore per spiegare come quark con caratteristiche identiche potessero coesistere all'interno degli adroni e al contempo soddisfare il principio di esclusione di Pauli, essendo i quark dei fermioni.
La carica di colore svolge un ruolo analogo a quello della carica elettrica in elettrodinamica, con la differenza che mentre esiste un solo tipo di carica elettrica, nel senso che un solo scalare è sufficiente per caratterizzare la carica di una particella, in QCD ci sono tre tipi di colore (rosso, verde e blu). In un processo di interazione forte, il colore del quark, ma non il suo sapore, può cambiare. Ad esempio, un quark up blu può convertirsi in un quark up rosso; poiché il colore è sempre conservato, ciò significa che il gluone deve portare una differenza, in questo caso, un'unità di blu e meno un'unità di rosso.
I gluoni, quindi, sono "bicolorati", nel senso che trasportano due cariche di colore, a differenza dei quark che ne trasportano solo una. Ci sono $3 \cdot 3 = 9$ possibilità considerando che ci sono tre colori, quindi ci si aspetterebbe che ci siano 9 tipi di gluoni, ma una di queste viene eliminata per delle considerazioni di simmetria, quindi i gluoni sono 8.
Un'altra differenza tra cromodinamica ed elettrodinamica è il valore della costante di accoppiamento. Come detto prima, ogni vertice di un diagramma di Feynman, per l'interazione elettromagnetica, introduce un fattore di $\alpha = \frac{1}{137}$. Sperimentalmente, la costante di accoppiamento corrispondente per le forze forti α_s ha un ordine di grandezza dell'unità, questo comporta che i diagrammi di Feynman ad ordini successivi contribuiscono sempre di più e la descrizione perturbativa, che ha funzionato così bene in QED, sembrava presentare un problema di divergenza nel caso della QCD.
Uno dei grandi successi della QCD è stata la scoperta che la quantità che svolge il ruolo di costante di accoppiamento in realtà non è affatto costante, ma dipende dalla distanza di separazione tra le particelle interagenti. Sebbene alle distanze relativamente grandi caratteristiche della fisica nucleare questa sia grande, a distanze molto brevi, ovvero al di sotto delle dimensioni lineari del protone, si riduce. Questo fenomeno è noto come *libertà asintotica*: significa che all'interno di un adrone, i quark si muovono senza interagire molto, comportamento che è stato riscontrato negli esperimenti di diffusione anelastica profonda. Da un punto di vista teorico, la scoperta della libertà asintotica ha salvato il calcolo di Feynman come strumento legittimo per la QCD, nel regime ad alta energia.
Come accennato in precedenza, i quark e i gluoni si combinano in modo che, qualsiasi adrone compongano, esso sia complessivamente neutro in termini di carica di colore. Sperimentalmente infatti, si osserva che i quark sono confinati in stati legati incolori di due (mesoni) e tre (barioni), sebbene di recente siano stati osservati anche adroni esotici formati da quattro quark (pentaquark). Questo fenomeno è chiamato *confinamento di colore*. Di conseguenza, quando i quark vengono prodotti negli acceleratori di particelle, invece di vedere i singoli quark

nei rivelatori, si rivelano *jet* di varie particelle neutre dal punto di vista della carica di colore (mesoni e barioni) raggruppate insieme. Questo processo è chiamato *adronizzazione*.

1.2.3 Interazione debole

Tutti i quark e leptoni interagiscono tramite interazione debole. I leptoni non hanno colore, quindi non partecipano alle interazioni forti; i neutrini non hanno carica, quindi non sperimentano forze elettromagnetiche; ma tutti si uniscono nelle interazioni deboli. Esistono due tipi di interazioni deboli: cariche, mediate dai bosoni $W^{\pm}$, e neutre, mediate dal bosone Z. Questi sono bosoni massivi aventi rispettivamente massa di circa $80\ Gev/c^2$ e $91\ Gev/c^2$, ciò rende la loro vita media dell'ordine di $10^{-24}\ s$ e tale aspetto limita il raggio d'azione dell'interazione debole, che risulta dell'ordine di $10^{-18}\ m$ (circa mille volte più piccolo del diametro del nucleo atomico).
Per quanto riguarda l'interazione debole carica, un esempio comune è un leptone a carica negativa che si converte nel corrispondente neutrino, con emissione di un W^- (o assorbimento di un W^+). Per esempio, il processo del decadimento del muone $\mu^- \longrightarrow e^- + \bar{\nu}_e + \nu_\mu$ è rappresentabile come in figura 1.3.

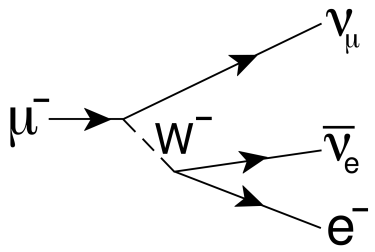


Figura 1.3: Diagramma di Feynman del decadimento del muone.

Similmente, un quark di tipo down (d, con carica $-\frac{1}{3}$) può essere convertito in un quark di tipo up (u, con carica $+\frac{2}{3}$), emettendo un bosone W^- o assorbendo un bosone W^+. C'è però una sostanziale differenza tra i due preocessi: la costante di accoppiamento debole è universale per tutti i leptoni, per i quark invece la probabilità calcolata nel vertice di interazione dipende dalla tipologia di quark interessati nel cambiamento di sapore. Nello SM la matrice Cabibbo-Kobayashi-Maskawa (matrice CKM) è una matrice unitaria che contiene informazioni sui decadimenti deboli con cambiamento di sapore. L'idea essenziale è che le generazioni di quark siano "distorte", ai fini delle interazioni deboli. Infatti, anziché avere gli autostati dell'interazione forte:

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

Quelli dell'interazione debole saranno:

$$\begin{pmatrix} u \\ d' \end{pmatrix} \begin{pmatrix} c \\ s' \end{pmatrix} \begin{pmatrix} t \\ b' \end{pmatrix}$$

dove d', s' e b' sono combinazioni lineari dei quark d, s e b:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}$$

Sperimentalmente, i moduli degli elementi della matrice (che è una matrice complessa) sono:

$$\begin{pmatrix} 0,97383 & 0,2272 & 3,96 \cdot 10^{-3} \\ 0,2271 & 0,97296 & 42,21 \cdot 10^{-3} \\ 8,14 \cdot 10^{-3} & 41,61 \cdot 10^{-3} & 0,999100 \end{pmatrix}$$

Osservando queste quantità si deduce che reazioni di tipo $u \leftrightarrow d$, $s \leftrightarrow c$ e $t \leftrightarrow b$ sono le più probabili.

1.2.4 Unificazione elettrodebole

Il formalismo vero e proprio dell'interazione debole nacque quando si capì che i processi di corrente neutra possono essere unificati con i processi elettromagnetici. La teoria di Glashow, Weinberg e Salam dell'unificazione elettrodebole (teoria GWS) inizia con quattro mediatori senza massa, ma è stata successivamente estesa con l'inclusione dal cosiddetto meccanismo di Higgs, che consente di attribuire massa alle $W^{\pm}$ e la Z, mentre uno rimane senza massa: il fotone γ. Sebbene sperimentalmente una reazione mediata da $W^{\pm}$ o Z sia abbastanza diversa da quella mediata da γ, secondo la teoria GWS sono tutte manifestazioni di una singola interazione elettrodebole. La relativa debolezza della forza debole è attribuibile alla massa dei bosoni vettoriali; la sua forza intrinseca è infatti leggermente maggiore di quella della forza elettromagnetica.
Nell'ambito di possibili scenari ad energie più elevate, teorie come la *Grand Unified Theory* (GUT) cercano un regime a cui tutte le interazioni fondamentali, ed in particolare le loro costanti di accoppiamento, possano convergere. Tale unificazione, se si realizza, è in genere prevista avvenire a scale molto maggiori delle energie caratteristiche del MS. Dalla forma funzionale delle costanti di accoppiamento correnti è possibile stimare l'energia alla quale avviene questa unificazione:

intorno a $10^{16}\ GeV$. Questa è, ovviamente, più alta di qualsiasi energia attualmente accessibile. Tale idea spiegherebbe la differenza di intensità tra le tre interazioni come derivante dal fatto che sono state finora osservate a scale di energia molto più basse della GUT, a cui invece sono espressione della stessa forza.

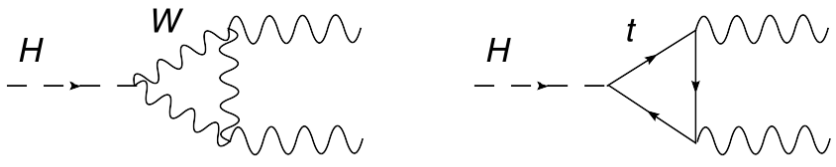
1.2.5 Il bosone di Higgs e i suoi decadimenti

Lo SM non è in grado di prevedere la massa del bosone di Higgs [2], tuttavia fornisce previsioni precise per quanto riguarda la sezione d'urto, grandezza che quantifica la probabilità di interazione delle particelle coinvolte in uno specifico processo, e il *branching ratio* (BR), legato invece alla frequenza con cui si verifica un processo. Con la scoperta del bosone di Higgs, ad una massa di circa 125 GeV, è stato anche possibile studiare nel dettaglio le altre sue proprietà.
Dallo SM sappiamo che il bosone di Higgs può decadere in coppie di fermioni o bosoni e i valori del BR, che sono legati alle costanti di decadimento parziali riferite ad uno specifico canale, sono riportati nella tabella 1.4.

Canale	Branching ratio
$H \to \gamma\gamma$	$2,28 \cdot 10^{-3}$
$H \to ZZ$	$2,64 \cdot 10^{-2}$
$H \to W^+W^-$	$2,15 \cdot 10^{-1}$
$H \to \tau^+\tau^-$	$6,32 \cdot 10^{-2}$
$H \to b\bar{b}$	$5,77 \cdot 10^{-1}$
$H \to Z\gamma$	$1,54 \cdot 10^{-3}$
$H \to \mu^+\mu^-$	$2,19 \cdot 10^{-4}$

Tabella 1.4: Branching ratio dei canali di decadimento del bosone di Higgs.

I decadimenti più probabili dunque sono $H \to b\bar{b}$ e $H \to W^+W^-$, che tuttavia sono di difficile rilevazione da un punto di vista sperimentale, non essendo ben distinguibili dal fondo, contrariamente a $H \to \gamma\gamma$ che, nonostante abbia un BR relativamente basso, è caratterizzato da uno stato finale in cui vi sono due fotoni sono molto energetici, quindi ben identificabili dagli attuali rivelatori. Infatti, tale canale è stato tenuto in considerazione come uno degli obiettivi più importanti durante la fase di progettazione del calorimetro elettromagnetico dell'esperimento CMS, che verrà approfondito nei prossimi capitoli. Il bosone di Higgs può decadere in due fotoni mediante loop di particelle cariche massive, e i due contributi principali sono dati da loop di bosoni W e di quark top.



(a) (b)

Figura 1.4: Contributi principali al decadimento del bosone di Higgs in due fotoni: (a) loop con il bosone W; (b) loop con il quark top.

1.3 Fisica oltre il Modello Standard

Lo SM funziona bene in un range di energie che va da pochi MeV alla scala TeV; ci sono, tuttavia, come già accennato, oltre all'esclusione dell'interazione gravitazionale, alcune osservazioni che non sono previste dalla teoria. La più evidente è dovuta al fatto che gli ingredienti di questo modello costituiscono circa il 5% della massa-energia totale dell'Universo. L'idea è che ci sia un altro tipo di materia, chiamata *materia oscura*, la cui prima evidenza deriva dall'osservazione delle curve di rotazione della galassia, dove la velocità di rivoluzione non segue la previsione delle teorie gravitazionali se la massa della galassia fosse solo quella degli oggetti conosciuti. Un'altra incongruenza con lo SM emerge studiando i neutrini provenienti da sorgenti diverse (dal Sole, dai reattori nucleari o dall'interazione dei raggi cosmici con l'atmosfera); è stato dimostrato che il loro sapore può cambiare mentre viaggiano dalla sorgente al rivelatore. Questo processo è noto come **oscillazioni dei neutrini** e coinvolge tutte e tre le famiglie leptoniche. La spiegazione fisica dell'oscillazione dei neutrini si basa sulla sovrapposizione di stati di massa, che è in netto contrasto con la previsione dello SM dei neutrini privi di massa.
Oggi sappiamo che la massa del bosone di Higgs è di circa 125 GeV. Da un punto di vista teorico, questo valore è influenzato dalla presenza di correzioni quantistiche da cui risultano valori molto più grandi di quello misurato. Avere una massa così piccola per il bosone di Higgs rispetto all'entità delle correzioni comporta una regolazione dei parametri dello SM per annullare quasi completamente il contributo di queste correzioni di ordine superiore. Questo problema è generalmente indicato come **problema di gerarchia**.
Nel corso degli anni sono stati sviluppati molti modelli che includono una spiegazione a questi fenomeni esclusi dallo SM, che ricadono nell'insieme di teorie *Beyond Standard Model* (BSM). Al momento, non sono state trovate prove che esulino dalla fisica del Modello Standard.

1.3.1 Vector Like Quark

La chiralità è una proprietà che distingue i sistemi fisici in destrorsi e sinistrorsi: un sistema fisico possiede una chiralità se sotto una trasformazione di parità si trasforma nel sistema con la chiralità opposta. Un fermione è definito *vector-like* se le sue chiralità sinistrorsa e destrorsa appartengono alla stessa rappresentazione del gruppo di simmetria $SU(2) \times U(1) \times SU(3)$ nello SM.
Sebbene al momento non ci siano prove dell'esistenza di quark detti vector-like [3] (**Vector like quark, VLQ**), da un punto di vista teorico, i VLQ vengono utilizzati in molti modelli differenti BSM e c'è una vasta letteratura sulle loro proprietà e fenomenologia. La presenza di correnti neutre che cambiano sapore sembrano essere una caratteristica distintiva dei VLQ. Diversamente dai quark, i VLQ hanno quattro sapori: **T**, **X**, **B** ed **Y** e possono trovarsi in configurazione di singoletto, doppietto o tripletto.
I principali canali di decadimento in particelle dello SM sono:

$$T \to W_+ b, Zt, Ht$$

$$B \to W_- t, Zb, Hb$$

$$B \to W_+ t$$

$$B \to W_- b$$

Lo studio di questi oggetti sono condotti agli esperimenti CMS e ATLAS di LHC durante la Run II con un'energia di centro di massa di $\sqrt{s} = 13$ TeV, in virtù del fatto che si presume abbiano una massa dell'ordine dei qualche TeV.
Un esempio di processo di produzione di VLQ singolo che decade in un quark top e un bosone di Higgs, il quale decade poi in due fotoni, è mostrato in figura 1.5. Tale processo sarà oggetto di studio nell'ultimo capitolo del presente lavoro di tesi.

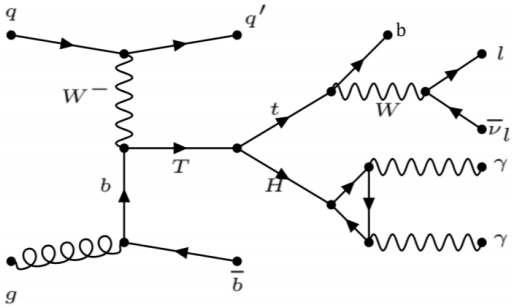


Figura 1.5: Diagramma di Feynman della produzione di un T e il suo canale di decadimento $T \to Ht$.

Capitolo 2

LHC e l'esperimento CMS

Negli anni '90, era teorizzata la produzione del bosone di Higgs nella collisione protone-protone, oltre che i suoi differenti modi di decadere, ed alcuni di questi erano strettamente legati alla massa - sconosciuta - della particella. Quindi, gli acceleratori che avrebbero avuto come obiettivo la ricerca del bosone di Higgs, dovevano coprire un intervallo di massa più ampio possibile. Il primo acceleratore a raggiungere la regione di massa in cui si sospettava potesse trovarsi il bosone di Higgs è stato il Large Electron-Positron Collider (LEP) del CERN, attivo dal 1989 al 2000. Questo acceleratore ha dato un gran contributo all'avanzamento della ricerca sulla particella stabilendo che la sua massa dovesse essere maggiore di 115 GeV. Ciò ha indirizzato la ricerca verso regimi energetici più elevati, contribuendo alla decisione di realizzare una macchina dedicata come il Large Hadron Collider.

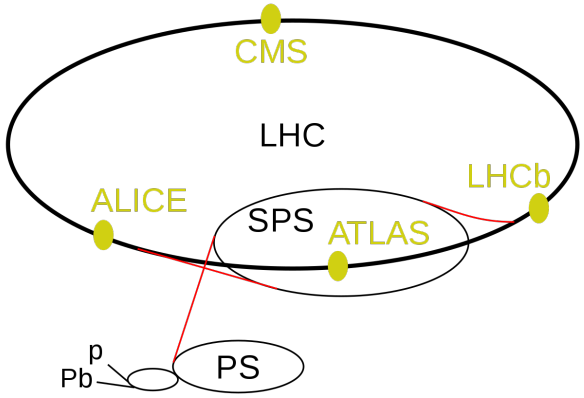
2.1 Large Hadron Collider

Inaugurato nel 2008 al CERN di Ginevra, Large Hadron Collider (LHC) [4] è un sincrotrone costruito in un tunnel sotterraneo lungo 27 km a circa un centinaio di metri sotto terra (un minimo di 50 m e un massimo di 175 m): il più grande e potente esistente attualmente. Responsabile della rilevazione del bosone di Higgs, LHC è un progetto in continua crescita che vede tra gli altri suoi obiettivi lo studio dello SM e la ricerca nuova fisica.

2.1.1 Caratteristiche

La macchina accelera due fasci di protoni (o alternativamente di ioni di Piombo) che circolano in direzioni opposte, all'interno dei tubi a vuoto, per poi scontrarsi in più punti del percorso. Prima di essere iniettati in LHC, i fasci di protoni da accelerare vengono 'strappati' dall'idrogeno gassoso per poi entrare in una serie

di acceleratori preliminari.
Il primo acceleratore che incontrano è **LINAC 2**, acceleratore lineare che con i suoi 36 m di lunghezza permette alle particelle di raggiungere energie fino a 50 MeV. Gli incontri successivi saranno con tre acceleratori circolari, chiamati sincrotroni, dal diametro via via maggiore. Il primo è il **Proton Synchrotron Booster**, con circa 50 m di diametro, raggiungendo così circa 1,5 GeV di energia. Immediatamente successivo, il **Proton Synchrotron** (**PS**) con 629 m di diametro, fa arrivare il fascio a 25 GeV. Dopo ancora, il **Super Proton Synchrotron** (**SPS**), di quasi 7 km di diametro, porta l'energia a 450 GeV.
È chiaro quindi che quanto più esteso è un acceleratore, tanto più le particelle al suo interno saranno energetiche e tanto più sarà alta l'energia del centro di massa relativa all'urto. Nel caso di due particelle che si scontrano frontalmente con uguali energie $E_1 = E_2 \equiv E$, l'energia disponibile per la creazione di nuove particelle è massima, in quanto è possibile produrre particelle con massa invariante $\sqrt{s} = 2E$.



Quando infine i fasci raggiungono LHC, vengono accelerati fino a raggiungere velocità di pochissimo inferiori alla velocità della luce, guidati su ben definite orbite da intensi campi magnetici generati da magneti a dipolo di 15 metri di lunghezza, che curvano i fasci generando un campo magnetico di 8 Tesla, e magneti a quadrupolo, ciascuno di 5-7 metri di lunghezza, che focalizzano i fasci. Appena prima della collisione, vengono utilizzati altri tipi di magneti per "spremere" (*squeezing*) i due fasci e aumentare le possibilità di collisioni, per un totale di più di 9000 magneti.
Per mantenere queste prestazioni, in questa macchina crioscopica troveremo temperature di circa 2 K (temperatura dell'elio liquido che garantisce la superconduttività dei magneti) e pressioni dell'ordine di 10^{-19} bar per rendere trascurabile la probabilità che i protoni accelerati collidano con le molecole di gas residuo all'interno dell'anello, in un sistema di vuoto artificiale più spinto mai stato realizzato.

Per avere un'idea precisa sull'effettive prestazioni di questo apparato, si tiene conto di un parametro specifico detto **luminosità**. Conosciuta la sezione d'urto σ e il *rate* $R = \frac{dN}{dt}$, ovvero il numero di collisioni per unità di tempo, la luminosità istantanea sarà data da:

$$L = \frac{1}{\sigma}\frac{dN}{dt} = \frac{R}{\sigma}$$

Ad una maggiore luminosità corrisponde quindi una maggiore probabilità di osservare fenomeni rari con σ più basse.
Se si integra la luminosità istantanea in un intervallo di tempo $\triangle T$ si ottiene la luminosità integrata, che dà il numero di eventi con una certa sezione d'urto:

$$L_{int} = \int_{\triangle T} L dt$$

che ha le dimensioni dell'inverso di una superficie e solitamente viene misurata in barn inverso (b^{-1}). La luminosità integrata totale del cosiddetto Run-II di LHC (2015-2018) è stata di circa 150 fb^{-1}.
Una delle limitazioni più importanti alla luminosità deriva dall'inevitabile interazione dei fasci (*beam-beam effects*) che corrono fianco a fianco nell'acceleratore. Dunque per mantenere alta la luminosità ad energie sempre maggiori bisogna mantenere contenuti i parametri che la caratterizzano, sottoponendo LHC a periodiche manutenzioni di aggiornamento, che coincidono con i periodi di *long shutdown* (LS), come mostrato in figura 2.1.

2.1.2 Esperimenti ad LHC

I dati di queste collisioni sono registrati da più sistemi di rivelatori [5], ad oggi in tutto sono sette, posizionati lungo l'anello di LHC, in punti differenti, ognuno specializzato nella rilevazione di diversi fenomeni. I più grandi sono:

1. **Compact Muon Solenoid** (**CMS**): costruito attorno a un enorme magnete a solenoide, da cui prende il nome, è uno dei due più grandi esperimenti presenti ad LHC. Per ovvi motivi, gli sarà dedicato un paragrafo dettagliato in seguito;

2. **A Toroidal LHC Apparatus** (**ATLAS**): insieme a CMS, utilizza rivelatori generici per indagare sulla più vasta gamma possibile di fisica, i due rivelatori sono tuttavia distinti da differenti compromessi di fabbricazione dei detector al loro interno: avere due rivelatori progettati in modo indipendente è fondamentale per la conferma incrociata di qualsiasi nuova scoperta;

Figura 2.1: Storia e piani futuri di LHC dal 2011 al 2040.

3. **Large Hadron Collider-beauty** (**LHCb**): come suggerisce il nome, questo rivelatore è specializzato nella fisica del quark bottom, ma anche nella ricerca di nuove sorgenti di violazione della simmetria Carica-Parità (CP), da cui scaturisce l'asimmetria materia-antimateria;

4. **A Large Ion Collider Experiment** (**ALICE**): si propone di portare avanti uno studio ad ampio raggio delle particelle prodotte nella collisione di ioni pesanti (nuclei di piombo) alle energie ottenibili da LHC.

Gli esperimenti più piccoli sono TOTEM, LHCf (rispettivamente nei pressi di CMS e ATLAS), che si concentrano sui protoni o ioni pesanti che interagiscono senza incontrarsi frontalmente quando i raggi entrano in collisione. Infile, MoEDAL utilizza rivelatori distribuiti vicino a LHCb per cercare l'ipotetico monopolo magnetico.

2.2 Compact Muon Solenoid

Il Compact Muon Solenoid [6] è un eterogeneo sistema di rivelatori costruito intorno ad uno dei punti di collisione di LHC, progettato in 15 sezioni separate lungo la direzione del fascio che sono state costruite in modo da assicurare la possibilità di accedere singolarmente e con facilità alle sezioni da riparare o potenziare senza arrecare minimo disturbo al resto della struttura, donandogli una struttura "a cipolla" con una forma cilindrica, come è evidente in figura 2.2.

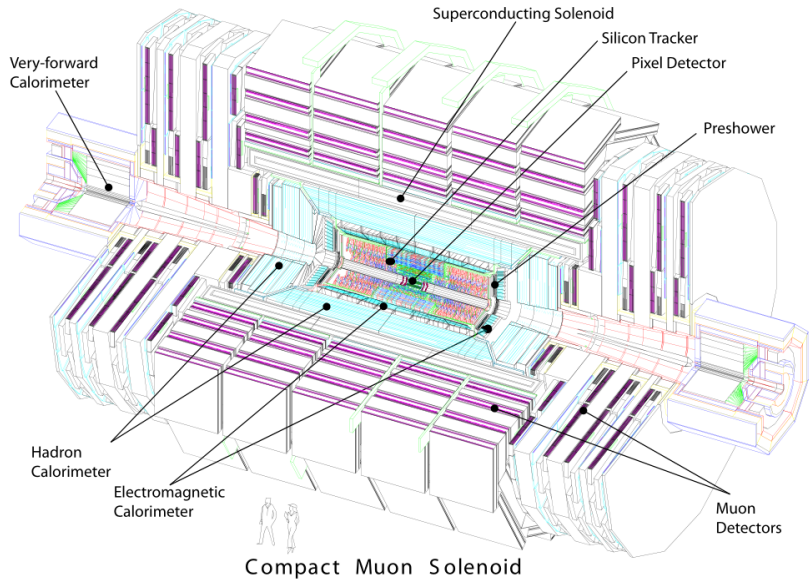


Figura 2.2: Schema della sezione del CMS.

2.2.1 Sistema di coordinate CMS

Data la sua forma, chiaramente è comodo adottare un sistema di coordinate cilindriche che ha l'origine centrata nel punto di collisione nominale all'interno dell'esperimento. Convenzionalmente la coordinata y punta verso l'alto, l'angolo azimutale Φ è misurato dall'asse x nel piano xy e la coordinata radiale su questo piano è indicata con R. L'angolo polare θ è misurato rispetto all'asse z. Una variabile correlata all'angolo con il quale le particelle cariche emergono dalle collisioni è chiamata *pseudorapidità* ed è definita come

$$\eta = -\ln \tan \frac{\theta}{2}$$

Quantificare questa grandezza è utile poiché $\triangle\eta$ risulta essere un invariante relativistica per *boost* lungo l'asse z. Con i parametri η e Φ è possibile inoltre costruire un'ulteriore invariante relativistica $\triangle R$ che rappresenta la distanza definita tra le posizioni nel piano (η, Φ) delle particelle interessate:

$$\triangle R = \sqrt{(\triangle\eta)^2 + (\triangle\Phi)^2}$$

Infine l'impulso di una particella nella rappresentazione cartesiana può essere scomposto nella componente longitudinale, lungo l'asse del fascio, che nel sistema di coordinate CMS è p_z, e nella componente trasversale, p_t, il cui modulo

$$p_t = \sqrt{p_x^2 + p_y^2}$$

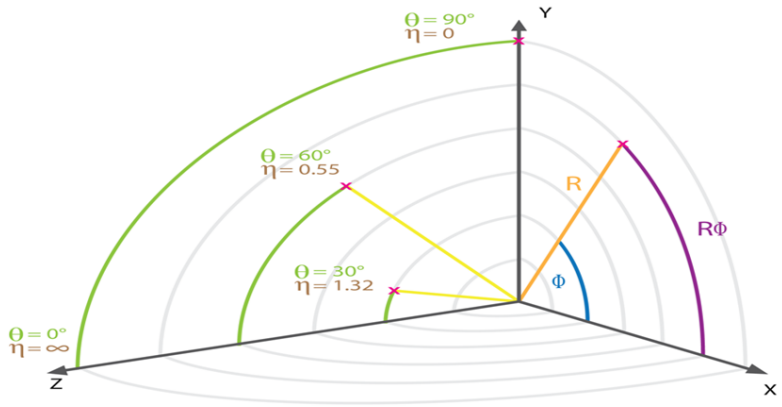


Figura 2.3: Sistema di riferimento utilizzato in CMS con la variabile η in funzione di alcuni angoli θ.

2.2.2 Struttura e sottorivelatori

Il sistema "a cipolla" di CMS è sostenuto da una struttura in ferro realizzata all'esterno dei sottorivelatori composta da 11 elementi di grandi dimensioni (il cui peso va da 400 tonnellate per il più leggero fino a 1920 tonnellate per la ruota centrale, che comprende il solenoide superconduttore e il suo criostato). Oltre ad una funzione strutturale, il giogo in ferro costituisce l'elemento di ritorno del campo magnetico. L'incavatura del magnete è abbastanza grande da contenere il tracker di silicio e i sistemi di calorimetri all'interno.

Tracker interno

Il primo sottorivelatore a ricevere informazioni sullo scontro dei fasci è il **sistema di tracciamento interno**, il tracker di silicio più grande mai costruito, che abbraccia il punto di interazione dei fasci e copre una regione di pseudorapidità $\eta = 2,5$. Esso si può suddividere in tre aree:

- la parte più interna, dove il flusso di particelle è più elevato, è composto da rivelatori a pixel (distribuiti in tre strati nel *barrel* ed in due negli *endcap*) con una risoluzione di $100 \times 150 \mu m^2$;
- regione intermedia, dove la diminuzione del flusso consente l'uso di microstrip al silicio con dimensione minima delle celle di $10cm \times 50\mu m$;
- regione esterna, dove il flusso è così basso da consentire l'uso di strisce di silicio più grandi, con dimensioni massime di $25cm \times 180\mu m$.

La funzione principale per cui è stato progettato è fornire una misurazione precisa ed efficiente delle traiettorie delle particelle cariche e il silicio viene utilizzato al fine di ottenere un'elevata granularità e una risposta rapida. Tuttavia, queste caratteristiche implicano un'elevata densità di potenza dell'elettronica del rivelatore che a sua volta richiede anch'esso raffreddamento. A questo proposito è stato necessario trovare un compromesso. L'intenso flusso di particelle causerà anche gravi danni da radiazioni al sistema di tracciamento, che sarà in grado di operare in questo ambiente ostile per una durata prevista di 10 anni. Incrociando le sue misurazioni con quelle del calorimetro elettromagnetico e del sistema dei muoni è possibile identificare gli elettroni e i muoni.

Calorimetro elettromagnetico

Il secondo sottorivelatore è il **calorimetro elettromagnetico** chiamato anche **ECAL**, composto da 61200 cristalli di piombo tungstato ($PbWO_4$, scelto poiché possiede un'alta resistenza alla radiazione) montati nella parte centrale, chiusi da 7324 cristalli in ciascuna delle due estremità. Anche ECAL può essere suddiviso in tre parti:

- Barrel EB, caratterizzato da $\eta < 1,48$;
- Endcap EE, dove $1,48 < \eta < 3$;
- Preshower, posto davanti agli endcaps, per cui $1,65 < \eta < 2,6$.

Un focus particolare va fatto su quest'ultimo, che è posizionato davanti ai cristalli di ECAL. Ha la forma di un disco spesso 20 cm la cui circonferenza è circa 2,5 m con un foro di 50 cm di diametro al centro. Esso è un prezioso strumento per la rilevazione del decadimento in due fotoni del bosone di Higgs; infatti, sebbene questo decadimento sia *gold-plated* (facile da vedere), ci sono alcuni tipi di eventi di fondo che possono imitare questo decadimento. Alcuni di questi fondi sono irriducibili, ad esempio due fotoni possono essere prodotti nella collisione protone-protone iniziale. Altri tipi di fondo sono, tuttavia, riducibili: se un pione neutro viene creato nella collisione protone-protone iniziale, decade quasi immediatamente in due fotoni ravvicinati, l'angolo tra questi due fotoni dipende dall'energia del pione e dalla sua direzione ed esso è molto più piccolo per i pioni. Il preshower ha una granularità molto più fine rispetto all'ECAL, con strisce rilevatrici larghe 2 mm, rispetto ai cristalli ECAL larghi 3 cm. Quando i fotoni apparentemente ad alta energia vengono trovati nell'ECAL, possiamo cercare le loro tracce nel preshower, aggiungendo l'energia depositata lì all'energia totale dall'ECAL per discernere se siano realmente singoli fotoni ad alta energia o coppie di fotoni.

Calorimetro adronico

Costruito in modo da circondare ECAL, il calorimetro adronico (HCAL) è l'addetto alla rilevazione di getti adronici (jets) e neutrini, o anche particelle esotiche con energia trasversale apparentemente mancante. Costruito in modo che si alternino strati di materiale assorbente (ottone) e materiale attivo (scintillatore), anch'esso tripartito in:

- Barrel (HB), con l'aggiunta di un ultriore calorimetro chiamato Outer Hadronic Calorimeter (HO), la cui pseudorapidità è $\eta < 1,3$;
- Endcap (HE) con $1,3 < \eta < 3$;
- Hadron Forward (HF), con $3 < \eta < 5$.

Risulta ristretto radialmente tra l'estensione esterna del calorimetro elettromagnetico (R = 1,77 m) e l'estensione interna del solenoide (R = 2,95 m). Ciò limita la quantità totale di materiale che può essere inserita per identificare i getti adronici; pertanto, un ulteriore calorimetro è posizionato all'esterno del solenoide in modo da essere complementare ad HCAL.

Magnete superconduttore

Immediatamente successivo all'HCAL c'è il **solenoide superconduttore**, elemento caratterizzante di CMS, che attraverso il campo magnetico di 3,8 T da lui generato, ha il compito di curvare la traiettoria delle particelle cariche in modo da, noto il raggio di curvatura, ricavarne la massa, la carica e l'impulso. Esso è lungo 12,5 metri con un diametro di 6 metri. Il materiale superconduttore sfruttato nel magnete è il niobio-titanio (NbTi), infatti il magnete funziona a 4,6 K per raggiungere il regime superconduttore del NbTi.

Rivelatore di muoni

Intervallato tra gli spazi vuoti del del giogo di ferro, il sistema di rilevazione dei muoni è bipartito in

- Barrel, con $\eta < 1,2$;
- Endcap, con $0,9 < \eta < 2,4$.

Il sistema per muoni utilizza tre tipi di rivelatori a gas per la loro identificazione. Il rilevamento del muone è un potente mezzo per riconoscere le tracce di processi interessanti rispetto all'elevatissimo tasso di background previsto all'LHC a regime di piena luminosità. Ad esempio, il decadimento del bosone di

Higgs in ZZ o ZZ*, che a loro volta decadono in 4 leptoni, nel caso in cui siano tutti muoni viene considerato come un altro decadimento *golden*. Oltre alla relativa facilità nel rilevare i muoni, la migliore risoluzione di massa di 4 particelle può essere ottenuta se tutti i leptoni sono muoni perché sono meno influenzati degli elettroni dalle perdite radiative nel materiale del tracker interno. Pertanto, come suggerisce il nome dell'esperimento, il rilevamento dei muoni è di fondamentale importanza per CMS: la misurazione precisa dei muoni è stata un tema centrale sin dalle sue prime fasi di progettazione.

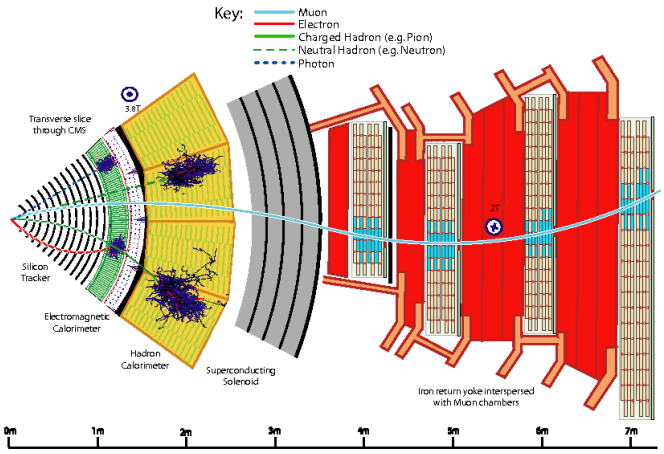


Figura 2.4: Percorsi delle particelle all'interno di CMS.

Sistema di trigger e acquisizione dati

Com'è facile immaginare, i tassi di collisione sono piuttosto elevati: per i protoni l'intervallo di attraversamento del fascio è di 25 ns, corrispondente a una frequenza di attraversamento 40 MHz. A seconda della luminosità, si verificano diverse collisioni ad ogni urto di fasci di protoni (circa 20 collisioni pp simultanee alla luminosità nominale di progetto di $10^{34}\ cm^{-2}s^{-1}$). È chiaro che sia impensabile memorizzare ed elaborare la grande quantità di dati associati all'elevato numero di eventi che ne risulta: è necessario ottenere una drastica riduzione della velocità di acquisizione. Questa attività viene eseguita dal sistema di trigger [7], che è l'inizio del processo di selezione degli eventi fisici. La velocità viene ridotta in due fasi chiamate trigger di livello 1 (L1) e trigger di alto livello (HLT), rispettivamente. La capacità di riduzione della velocità è progettata per essere almeno di un fattore 10^6 per il trigger L1 e HLT combinati.
In CMS tutti gli eventi che superano il trigger di Livello 1 (L1) vengono inviati a una computer farm che esegue selezioni utilizzando software di ricostruzione

per filtrare gli eventi. Esso è costretto a sostenere una velocità di uscita massima di 100 kHz, per un flusso di dati di circa 100 GByte/s. I candidati di L1 passano poi per HLT che è un'architettura software che mira a ridurre ulteriormente la frequenza degli eventi a circa 800 Hz. Gli eventi che passano l'HLT vengono quindi archiviati su disco locale pronti per essere ulteriormente elaborati. Per identificare le particelle al meglio è necessario correlare le informazioni derivanti da tutti gli strati del rivelatore. Questo approccio olistico è chiamato ricostruzione del *particle-flow* (PF) [8], da cui prende il nome l'algoritmo fondamentale per eseguire una rapida calibrazione incrociata dei vari sub-rivelatori, per convalidare le loro misurazioni e identificare i fondi di ognuno.

Capitolo 3

Ricostruzione dei decadimenti $H \to \gamma\gamma$ con il machine learning

Il presente lavoro di tesi si incentra sulla ricostruzione del bosone di Higgs attraverso l'analisi del canale di decadimento $H \to \gamma\gamma$, proveniente da processi in cui è prodotto in associazione ad un quark top: il processo di produzione associata nello SM (tHq), mostrata in figura 3.1, e la produzione attraverso l'ipotetico canale VLQ $T \to Ht$ rappresentati in 1.5.

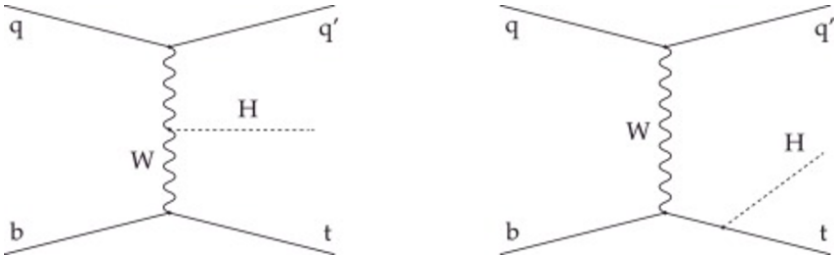


Figura 3.1: Diagrammi di Feynman dominanti per la produzione di eventi *tHq*.

La tipica analisi di fisica ad LHC consiste in una scelta di un insieme di richieste di selezione (dette "tagli") da applicare su alcune variabili a seguito delle collisioni protone-protone, al fine di separare il segnale dal fondo. Vista la grande mole di dati prodotti dall'acceleratore e acquisiti da CMS, si trae notevole vantaggio da un'analisi mediante tecniche di *Machine Learning* (ML), in modo da poter delegare ad un algoritmo automatizzato l'abilità di discernimento.

3.1 Introduzione al machine learning

Una definizione operativa di un algoritmo di apprendimento automatico può essere la seguente: "un programma che apprende dall'esperienza *E* con riferimento ad alcune classi di compiti *T* e con misurazione della performance *P*, se le sue performance nel compito *T*, come misurato da *P*, migliorano con l'esperienza *E*". La macchina quindi apprende se c'è un miglioramento delle prestazioni dopo il compito svolto, è dunque in grado di fare qualcosa di simile agli umani ragionamenti induttivi, generalizzando dalla propria esperienza. L'addestramento del sistema intelligente avviene su un insieme finito di dati, chiamato *training set*, e le sue prestazioni sono valutate su un ulteriore insieme di dati, chiamato *test set*.
I compiti del ML vengono classificati in tre paradigmi, a seconda della natura dei dati utilizzati o del feedback del sistema intelligente:

- **apprendimento supervisionato**: al modello vengono forniti degli esempi nella forma di possibili input e i rispettivi output desiderati (dati *etichettati*) e l'obiettivo è quello di estrarre una regola generale che associ l'input all'output corretto;
- **apprendimento non supervisionato**: il modello ha lo scopo di trovare una struttura negli input forniti, senza che gli input vengano etichettati;
- **apprendimento per rinforzo**: un agente intelligente interagisce in un ambiente dinamico nel quale cerca di raggiungere un obiettivo, la qualità di un'azione è data da un valore numerico di "ricompensa".

Per addestrare il modello, gli algoritmi automatici utilizzano una **funzione obiettivo** per misurare quanto bene il modello si adatta ai dati di addestramento. Dato un set di dati $\bar{\theta}$, essa è generalmente definita come:

$$obj(\theta) = L(\theta) + R(\theta)$$

dove *L* è chiamata *training loss*, che spesso viene scelta come uguale allo scarto quadratico medio, mentre *R* è il termine di regolarizzazione che tende a limitare la possibilità di un *overfitting*, fenomeno che riguarda una "memorizzazione" del training set che rende l'algoritmo non è in grado di generalizzare il modello sui dati del test set.

3.1.1 Decision tree - XGBoost

È stato utilizzato un algoritmo di tipo *decision tree*, cioè un modello predittivo rappresentabile come un grafo in cui ogni nodo descrive una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella caratteristica e

una foglia rappresenta il valore predetto per la variabile obiettivo a partire dai valori delle altre caratteristiche, che nel tree è rappresentato dal cammino dal *nodo radice* al *nodo foglia*. Esso è un algoritmo iterativo che ad ogni passo divide i dati finché questi non appartengono alla stessa classe in ciascun nodo. Questo processo, per evitare il rischio di overfitting, viene arrestato ad un fissato limite massimo di profondità del tree.

Il modello di *boosting* scelto per il tree per questo lavoro è **XGBoost** (XGB)

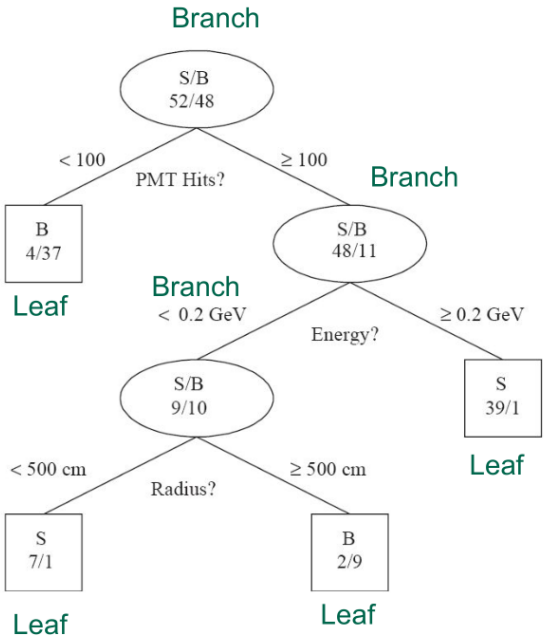


Figura 3.2: Esempio della struttura di un decision tree.

[9], acronimo di *eXtreme Gradient Boosting*. Esso è composto da un insieme di decision tree, in modo da considerare la somma degli output di ogni singolo tree che lo costituisce utilizzando un metodo chiamato *boosting*, che combina gli apprendimenti in sequenza in modo che ogni nuovo tree corregga gli errori di quello precedente, ciò rende l'algoritmo più performante e stabile che utilizzare un singolo tree. Dati K tree costituitivi, il modello può essere scritto:

$$y_i = \sum_{i=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

dove f_k è una funzione dello spazio funzionale $\mathcal{F}$ di tutti i possibili set di tree. Il valore predetto $\hat{y}_i^{(t)}$ al passo t è dato da:

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$
$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$
$$...$$
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Quindi, utilizzando come training loss lo scarto quadratico medio come esempio, essa può essere scritta come:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=0}^{t} R(f_i) =$$

$$= \sum_{i=1}^{n} l(y_i, (\hat{y}_i^{(t-1)} + f_t(x_i))) + R(f_t) + cost.$$

Sviluppando in serie di Taylor la *training loss* fino al secondo ordine:

$$obj^t = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + R(f_t) + cost.$$

dove g_i ed h_i sono definiti come:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

Dopo aver rimosso tutte le costanti, la funzione obiettivo al passo t sarà:

$$obj^t = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + R(f_t)$$

3.1.2 Configurazione di XGB per il riconoscimento dell'Higgs

XGBoost è un classificatore con una serie di variabili proprie, chiamate iperparametri, che ne configurano il funzionamento. Si riporta di seguito la tabella 3.1 con i valori utilizzati per gli iperparametri e il loro significato.

*La training loss utilizzata per questo lavoro è chiamata *logistic loss* ed è definita come:

$$\sum_{i=1}^{n} [y_i ln(1 + e^{-\hat{y}_i}) + (1 - \hat{y}_i) ln(1 + e^{\hat{y}_i})]$$

Iperparametro	Tipo	Valore	
number of estimators	integer	200	numero di tree da allenare
learning rate	float	0.1	velocità di apprendimento ad ogni passo
max depth	integer	4	profondità massima dell'tree
minimum child weight	integer	4	somma minima del peso dell'istanza
reg α	float	0.01	coefficiente di regolarizzazione
objective function	string	logloss*	funzione obiettivo utilizzata

Tabella 3.1: Valori degli iperparametri di XGB scelti per il presente lavoro di tesi

3.2 Test d'ipotesi

Per determinare le prestazioni dell'allenamento è stato definito un test d'ipotesi [10] basato sulla selezione effettuata dal classificatore. Essendo il decision tree soggetto all'errore, la nozione di errore statistico è parte integrante di tale test. Esso consiste nella scelta di due proposizioni concorrenti chiamate ipotesi nulla H_0 e ipotesi alternativa H_1. Si presume che l'ipotesi nulla sia vera fino a quando i dati non la smentiscano.
Per i problemi di classificazione, si utilizzano i termini positivo e negativo in riferimento alla previsione del classificatore, mentre i termini vero e falso si riferiscono al fatto che tale previsione corrisponda o meno alla realtà.

	H_0 vera	H_0 falsa
H_0 rigettata	falsi positivi (FN) errore di tipo I α	veri positivi (TP) inferenza corretta $(1-\beta)$
H_0 non rigettata	veri negativi (TP) inferenza corretta $(1-\alpha)$	falsi negativi (FN) errore di tipo II β

Tabella 3.2: Tanto più bassi saranno i falsi positivi e i falsi negativi, tanto più il test sarà valido.

Se il risultato del test corrisponde alla verità, è stata presa una decisione corretta. Tuttavia, ci sono due situazioni in cui la decisione è sbagliata: se si rifiuta H_0 che nella realtà è vera, allora si dice che si è commesso un errore di I tipo (FN); accettando invece H_0 falsa si commette un errore di II tipo (FP). Alla luce di ciò, possiamo rappresentare il segnale e il fondo in un grafico come in figura 3.3.

Un test è preferibile ad un altro se a parità di efficienza, anche chiamata tasso di veri positivi (TPR), si ha una probabilità di errore, anche chiamata tasso di falsi

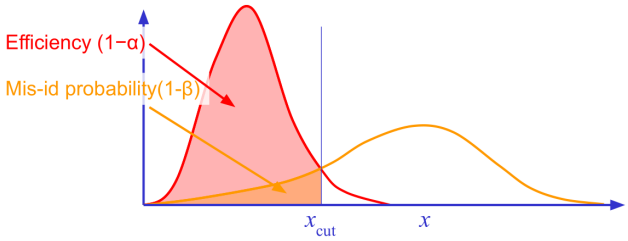


Figura 3.3: Per $x \leq x_{cut}$ si ha il segnale e per $x > x_{cut}$ si ha il fondo. I risultati ottenuti dal segnale si sovrappongono ai risultati ottenuti dal fondo. Variando il valore di x_{cut}, cambiano l'efficienza e la probabilità di errore.

positivi (FPR), più bassa. Tali indici si definiscono come:

$$TPR = \frac{\sum veri\ positivi}{attualmente\ positivi}$$

$$FPR = \frac{\sum falsi\ positivi}{attualmente\ negativi}$$

La curva TPR vs FPR è chiamata *Receiver Operating Characteristic* (ROC) ed è rappresentata in figura 3.4.
Valutando l'area sottesa alla curva ROC (AUC), è possibile avere un indicatore della capacità del classificatore di discernere. Il valore di AUC, compreso tra 0 e 1, è un buon indice per il confronto di modelli differenti.
Le curve ROC passano per i punti (0,0) e (1,1), hanno inoltre due condizioni che rappresentano due curve limite:

- una che taglia il grafico a 45°, passando per l'origine. Questa retta rappresenta il caso del classificatore casuale (*random choice*) e l'area sottesa AUC è pari a 0,5;
- la seconda curva è rappresentata dal segmento che dall'origine sale al punto (0,1) e da quello che congiunge il punto (0,1) a (1,1), avendo un'area sottesa di valore pari a 1, ovvero rappresenta il classificatore perfetto.

Solitamente, nel caso di algoritmi di classificazione binaria la cui risposta è compresa tra 0 e 1, è ragionevole porre $x_{cut} = 0.5$, misurando così le performances in termini di *recall* e *precision*. La prima equivale all'efficienza, $(1 - \alpha)$ oppure $(1 - \beta)$, e la seconda corrisponde a quante delle istanze selezionate sono vere.

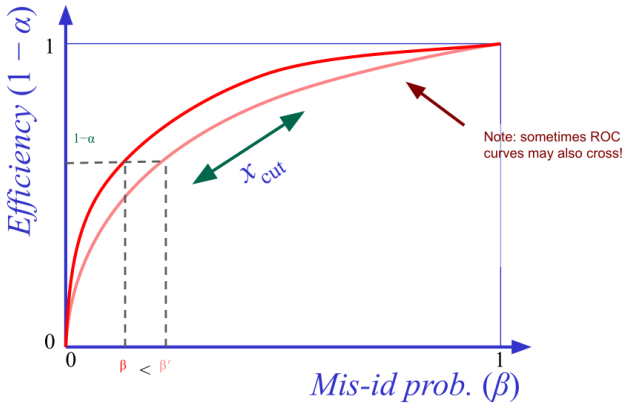


Figura 3.4: Esempio di due curve ROC a confronto.

3.3 Costruzione dei dati

L'analisi è stata eseguita su campioni ottenuti da simulazione Monte Carlo (MC) alle condizioni della presa dati a 13 TeV nel Run-II di LHC su segnali di tipo SM definiti in 1.2 e BSM realizzati entrambi con il software Madgraph. Il primo si riferisce ad una produzione associata di un quark top, un bosone di Higgs ed un quark spettatore aggiuntivo; il secondo si riferisce ad una produzione singola associata al vector like quark T rappresentata in 1.5, ipotizzando valori della massa di T pari a 600 GeV/c^2, 1000 GeV/c^2 e 1200 GeV/c^2.
Il decadimento oggetto di studio $H \to \gamma\gamma$ ha un BR piuttosto basso $(2,28 \cdot 10^{-3})$, dunque i file utilizzati hanno solo questo tipo di eventi al fine di facilitarne lo studio.

3.3.1 Rilevamento dei fotoni

Essendo i fotoni gli oggetti in esame da cui partirà la ricostruzione dei bosoni di Higgs, è opportuno descrivere la procedura del rilevamento di queste particelle da parte di CMS [11].
La presenza di tali particelle è rivelata principalmente in ECAL, dove sciami di elettroni e fotoni depositano la loro energia in diversi cristalli suddivisi in *cluster*. La presenza del calorimetro stesso e del materiale davanti al calorimetro si traduce in bremsstrahlung per gli elettroni e in conversioni di fotoni in coppie elettrone-positrone ($e^- e^+$). Sommando l'energia misurata in queste sezioni si ottengono le migliori prestazioni per i fotoni non convertiti o per gli elettroni. Per la corretta ricostruzione dell'energia diffusa è necessario sommare i contributi di tutti i cluster in quello che viene chiamato *supercluster*. L'aggiunta di energia depositata

nel preshower è la fase finale della procedura di ricostruzione ECAL. Il discernimento tra fotoni ed elettroni viene fatto incrociando i dati di ECAL con quelli del tracker interno, dove troveremo le tracce di particelle cariche. Sulla base di queste considerazioni, sono stati implementati numerosi algoritmi che ricostruiscano la particella in esame.

3.3.2 Ricostruzione del segnale e del fondo

Per una preliminare fase di costruzione del segnale e del fondo è stata utilizzata la piattaforma software *Root* [12], sviluppata nel linguaggio C++ dal CERN proprio per l'analisi dei dati in fisica delle alte energie.
I fotoni cercati, dal momento che non ci sono altre particelle nello stato finale del processo, dovranno essere due fotoni isolati, ovvero idealmente non accompagnati da altre particelle nell'evento.

Il criterio utilizzato per selezionare preliminarmente i fotoni è un taglio sul p_t,

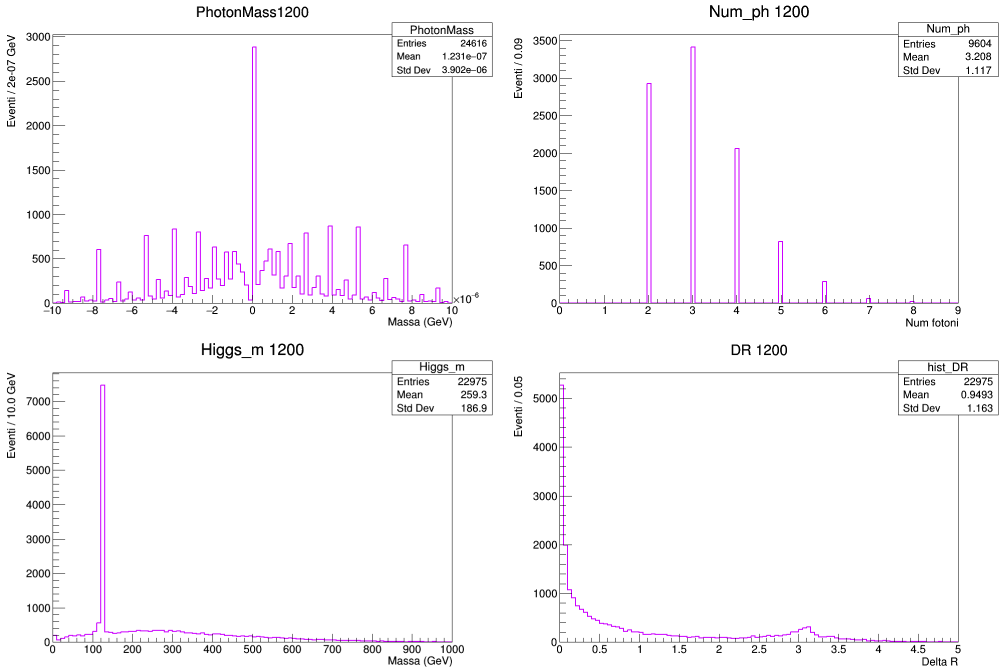


Figura 3.5: Caratteristiche estratte dal file BSM con massa di T pari a 1200 GeV:
- in alto a sinistra: distribuzione di massa dei fotoni;
- in alto a destra: numero di fotoni per evento;
- in basso a sinistra: distribuzione di massa degli Higgs ricostruiti dai fotoni;
- in basso a destra: distribuzione della distanza angolare $\triangle R$.

scartando i fotoni con valori di $p_t < 30\ GeV/c$. Per ognuno dei fotoni sopravvissuti a questa prima selezione, è stato calcolato il 4-vettore di Lorentz basato su un sistema di riferimento tale da avere come coordinate p_t, η, Φ ed M. Sommando a due a due i 4-vettori di Lorentz dei fotoni presenti per ogni evento, è stata ottenuta la popolazione dei bosoni degli Higgs candidati, descritti da variabili cinematiche quali la massa, il p_t, η e Φ.
L'obiettivo è creare l'input da dare ad un algoritmo supervisionato di classificazione binaria, ovvero volto a stabilire a quale di due categorie appartiene un'istanza di dati. L'input dell'algoritmo di classificazione è quindi il set di fotoni generato con annesse caratteristiche (tabulate in 3.3) e corrispondenti etichette (*truth* nel gergo) in cui ogni etichetta è un numero intero 0 o 1, utilizzato per discernere gli Higgs veri da quelli falsi.

Caratteristica	Tipo	Descrizione
eta	continua	η di entrambi i fotoni
phi	continua	Φ di entrambi i fotoni
mass	continua	massa di entrambi i fotoni
pt	continua	p_t di entrambi i fotoni
electronVeto	binaria	fotone vero oppure elettrone
hoe	continua	probabilità che sia un fotone vero o un elettrone
mvaID	binaria	ID derivante da un'analisi multivariata

Tabella 3.3: Caratteristiche dei fotoni utilizzate per la costruzione del segnale e del fondo.

Per creare la variabile *truth* che contraddistingue il segnale sono stati esclusi tutti i fotoni spazialmente lontani, quindi è stato imposto che i fotoni cercati dovessero essere ad una distanza angolare $\triangle R < 0.1$ e in più è stata utilizzata una variabile booleana la quale fa riferimento ad informazioni a livello di simulazione della particella ricostruita, che permette di separare il fotone da altre particelle cariche ricostruite male.

Dai grafici della figura 3.6 risulta evidente che il p_t del bosone di Higgs derivante dal decadimento di T ha una distribuzione con un picco, come ci si aspetta, intorno a valori di p_t che sono circa la metà della massa del T considerato. Essendo i VLQ T particelle piuttosto massive, si avranno quindi, come caratteristiche degli Higgs derivanti dai dati BSM, valori di p_t più alti rispetto ai dati SM.

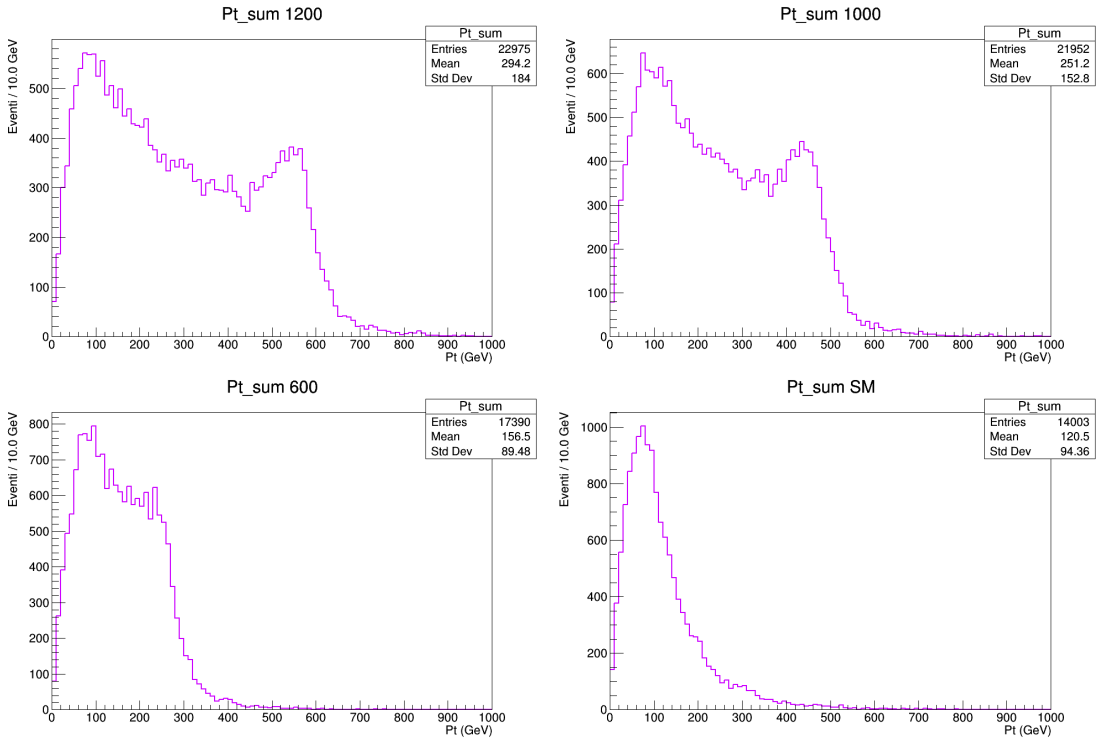


Figura 3.6: Distribuzioni del p_t del bosone per le seguenti configurazioni:
- in alto a sinistra: massa di T = 1200 GeV;
- in alto a destra: massa di T = 1000 GeV;
- in basso a sinistra: massa di T = 600 GeV;
- in basso a destra: SM.

3.4 Risultati dell'allenamento: coppie di fotoni

Una volta formattati i dati, il passo successivo è stato allenare il classificatore XGBoost. A questo scopo, è stato scelto il pacchetto Python *xgboost* [13] e il pacchetto *pandas* [14] per la gestione dei dati e l'interfaccia con il classificatore. La definizione del training set e del test set è stata fatta in due modi:

- separando casualmente i dati relativi ad ogni singola simulazione MC in modo da avere che il 70% costituisca il training set e il restante 30% sia utilizzato come test set, per testare che i risultati non dipendano statisticamente dal campione;
- definendo configurazioni miste allenando e poi testando su due set di dati derivanti da simulazioni MC differenti, per cercare eventuali bias dovuti al campione;

Delle caratteristiche cinematiche η e ϕ è stato preferibile calcolarne la somma e la differenza per ogni coppia di fotoni anziché mantenere tali caratteristiche legate al singolo fotone.
Per ogni oggetto fisico solitamente viene definita una variabile ID, cioè una variabile di identificazione derivante da un'analisi multivariata, in base alla quale sono definiti diversi livelli di selezione, detti "punti di lavoro" (WP, acronimo di *Working Point*) scelti in base all'efficienza di selezione.
Per i fotoni candidati sono state utilizzate le variabili *mvaID WP90* e *mvaID WP80*, che sono ID booleane con punti di lavoro rispettivamente al 90% ed all' 80% di efficienza.
Per ogni configurazione *train-test*, il classificatore è stato allenato sia utilizzando *mvaID WP90* e *mvaID WP80* come caratteristiche, insieme a quelle tabellate in 3.3, sia filtrando i dati prima dell'allenamento con queste stesse ID, in modo da confrontare le prestazioni dei tre allenamenti. Sono stati etichettati in tabella 3.4 rispettivamente come *no sel.*, *WP90* e *WP80*, dove sono stati riportati i valori delle AUC in percentuale per tutte le configurazioni considerate.

Nelle configurazioni split 0.7 train / 0.3 test le AUC sono leggermente più alte, questo è atteso in quanto i dati del training set e del test set riguardano fotoni che provengono dal medesimo fenomeno fisico. Invece, i risultati delle AUC riferiti alle configurazioni miste ci evidenziano dei bias da ricondurre probabilmente al p_t dell'Higgs che, come è evidente nella figura 3.6, è una caratteristica visibilmente discriminante tra i diversi fenomeni che fanno da genesi al bosone di Higgs.
Confrontando i risultati sembra evidente un calo delle prestazioni del classificatore quando questo viene allenato su un set di dati sopravvissuto dopo la selezione WP90 o WP80: ciò dimostra che è preferibile utilizzare dei dati non filtrati, permettendo così al classificatore di non perdere generalità. Infatti, effettuando questi tagli, nonostante sopravvivano i candidati migliori, si esclude al contempo una

train - test	no sel.	WP90	WP80
1200 - 1200	96.3	88.1	90.3
1000 - 1000	96.3	88.3	88.3
600 - 600	94.7	85.7	86.9
SM - SM	92.9	83.6	84.6
1200 - 1000	96.1	88.5	87.7
1200 - 600	92.7	85.2	85.1
1200 - SM	87.8	76.7	77.6
1000 - 1200	96.4	88.6	87.7
1000 - 600	93.7	86.2	86.2
1000 - SM	89.3	81.6	81.4
600 - 1200	94.3	86.5	86.1
600 - 1000	94.6	87.3	86.7
600 - SM	91.6	83.8	83.2
SM - 1200	95.7	88.0	87.3
SM - 1000	95.4	86.9	86.3
SM - 600	94.7	85.8	85.7

Tabella 3.4: AUC in percentuale delle configurazioni *split* 70% train / 30% test e delle configurazioni miste.

considerevole fetta di informazione necessaria al ML per il discernimento. Inoltre, dalle AUC non si evidenzia una grande differenza tra i filtri WP80 e WP90. Considerando che la popolazione di Higgs candidati che sopravvive alle selezioni è sicuramente minore di quella di partenza, è possibile, per avere un'idea quantitativa di tali filtri, definire un'efficienza di segnale e una di fondo come:

$$E_s = \frac{TP_{WP\ selection}}{TP_{full\ training}} \qquad E_f = \frac{TN_{WP\ selection}}{TN_{full\ training}}$$

In modo da effettuare delle correzioni sui TPR e i FPR:

$$TPR^{corr}_{WP\ selection} = E_s \cdot TPR_{WP\ selection} \qquad FPR^{corr}_{WP\ selection} = E_f \cdot FPR_{WP\ selection}$$

Inoltre, per un'analisi più approfondita e dunque un confronto migliore, sono stati definiti dei punti di lavoro riportando l'efficienza fissato il FPR a 10%, 1% e $0,1\%$ dell'allenamento, in modo da confrontare le velocità di apprendimento dell'algoritmo sui dati con e senza preselezione.

Dai dati tabulati in 3.5 (si noti che per le configurazioni *1200 train* $E_f < 0.01$ per la selezione WP80, dunque non è presente il punto di lavoro al 10%) è possibile osservare che tendenzialmente si hanno prestazioni migliori utilizzando un campione senza preselezione e che il filtro WP80 ha quasi sempre punti di lavoro più bassi rispetto agli altri allenamenti.

	no sel.			WP90			WP80		
train / test	10%	1%	0,1%	10%	1%	0,1%	10%	1%	0,1%
1200 - 1200	93.7	28.1	3.9	69.3	34.2	2.2	-	33.1	2.4
1000 - 1000	93.6	34.7	3.5	70.0	27.6	3.3	60.7	31.4	1.9
600 - 600	86.8	24.5	0.6	71.9	21.1	3.3	61.6	27.2	4.1
SM - SM	78.8	12.5	0.5	70.5	15.4	2.8	59.6	19.0	3.9
1200 - 1000	93.0	32.9	2.9	69.2	34.3	6.9	-	32.8	6.7
1200 - 600	75.7	18.0	1.9	68.8	29.8	1.8	-	28.8	4.2
1200 - SM	52.3	11.1	2.5	67.5	23.3	4.4	-	25.1	4.6
1000 - 1200	93.9	34.2	5.1	69.8	29.7	3.2	60.7	28.5	4.6
1000 - 600	80.5	20.4	3.1	60.6	22.5	3.9	60.6	27.9	3.0
1000 - SM	60.1	13.3	2.3	69.4	22.8	3.8	60.6	22.5	3.9
600 - 1200	88.4	24.5	2.9	71.6	20.1	1.6	61.4	20.1	1.9
600 - 1000	88.7	24.2	3.5	71.9	22.2	2.4	61.6	20.7	2.9
600 - SM	72.5	17.3	2.3	70.7	21.9	2.1	61.4	21.4	2.1
SM - 1200	91.9	28.8	1.9	70.5	19.9	3.1	60.2	18.9	3.1
SM - 1000	90.9	24.5	3.3	70.9	20.1	1.4	60.1	18.5	1.9
SM - 600	86.3	25.0	4.2	70.4	16.3	2.5	60.1	17.4	2.4

Tabella 3.5: Punti di lavoro in percentuale a 10%, 1% e 0,1% per configurazioni con e senza le preselezioni WP.

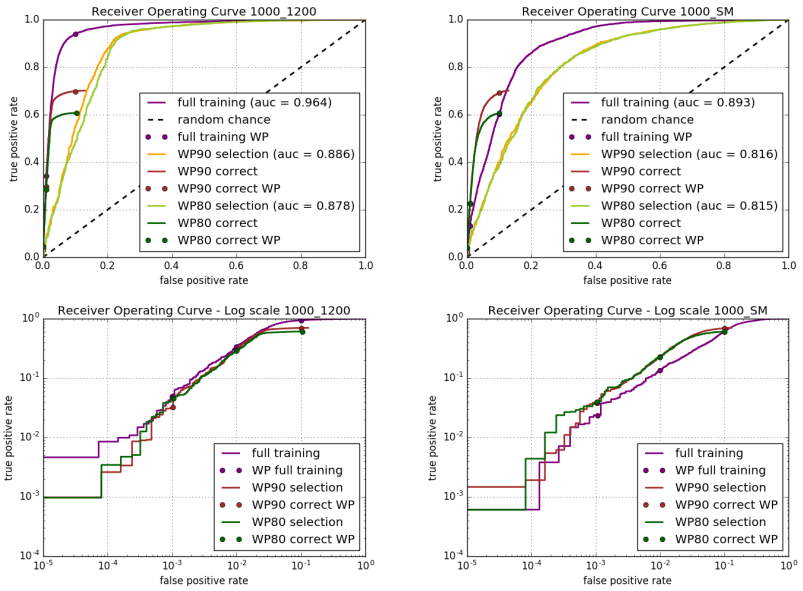


Figura 3.7: A sinistra: ROC *1000 test - 1200 train*; a destra: ROC *1000 test - SM train*, anche in scala logaritmica.

Per le configurazioni *1200 train - SM test* oppure *1000 train - SM test*, ove è presente una differenza sostanziale delle caratteristiche cinematiche (in primis del p_t) dei candidati Higgs appartenenti rispettivamente al training set e al test set, il punto di lavoro al 10% è considerevolmente più alto per la selezione WP90, rispetto a tutte le altre configurazioni, dove l'allenamento senza preselezione sembra essere quello più efficiente.
Per visualizzare i risultati sono state rappresentate le ROC per ogni configurazione, anche in scala logaritmica in modo da osservare al meglio i punti di lavoro. In figura 3.7 sono state riportate le ROC delle configurazioni ove si evidenziano maggiori differenze tra gli allenamenti con e senza preselezione.

È interessante notare, confrontando i punti di lavoro a 1% e $0,1\%$, che spesso le ROC si intersecano in vari punti, anche per configurazioni le cui AUC riportate nella tabella 3.5 suggeriscono un netto miglioramento per il caso senza preselezione, come mostrato in figura 3.8. Ciò significa che per indagare quale sia il metodo migliore per allenare il classificare non vi è una risposta assoluta per tutti i casi, ma è necessario fissare la probabilità d'errore e sceglierla in base alla reiezione del fondo e del regime cinematico dell'analisi.

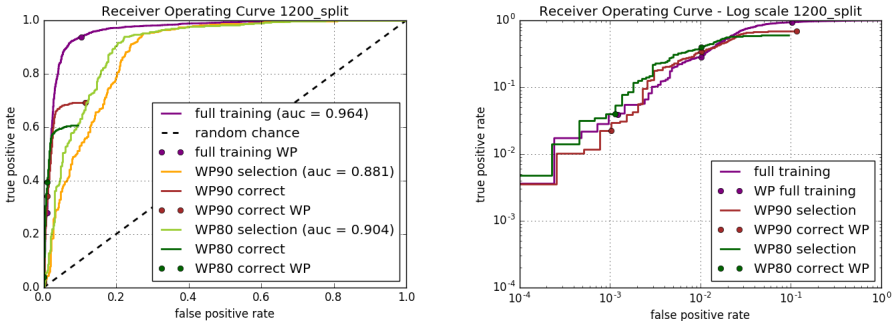


Figura 3.8: A sinistra: ROC configurazione *1200 split*; a destra: ROC in scala logaritmica.

I confronti tra le AUC e i punti di lavoro sembrano suggerire, come accennato in precedenza, un *bias* dell'allenamento legato all'impulso trasverso dei candidati Higgs. Ciò è osservabile nelle distribuzioni del fondo e del segnale delle configurazioni scelte riportate in figura 3.9, dove si nota che l'algoritmo riesce a discernere bene il segnale dal fondo in fase di training, ma quando testa ciò che ha imparato su un set di dati estrapolati da configurazioni di massa di T differenti, non riesce ad avere prestazioni ottimali, poiché *memorizza* le caratteristiche cinematiche del training test.

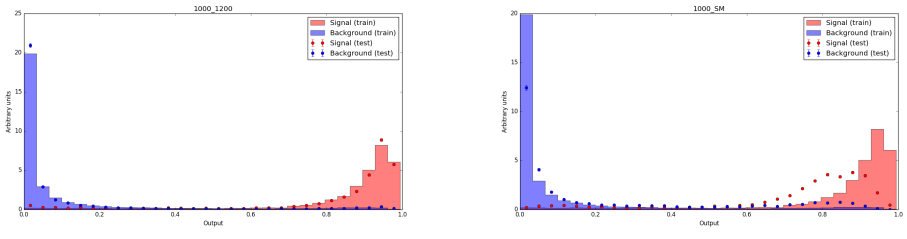


Figura 3.9: A sinistra: segnale e fondo della configurazione *1000 train - 1200 test*; a destra: segnale e fondo della configurazione *1000 train - SM test*.

Per indagare più a fondo l'importanza del p_t è possibile creare dei *rank* di caratteristiche di input in base a quanto queste siano informative per distinguere il segnale dal fondo, tramite tecniche che assegnano ad ogni variabile un punteggio chiamato *F-score*, descritto in [13] come somma di tutte le volte che il classificatore esegue una divisione sulla variabile considerata.

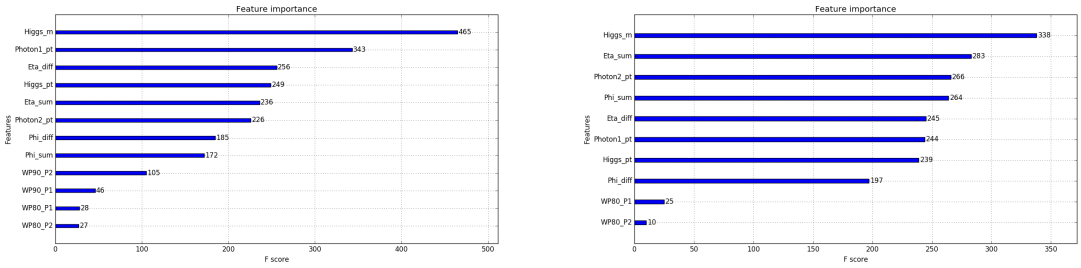


Figura 3.10: F-score delle caratteristiche di input per la configurazione di allenamento *1200 train - 600 test*. A sinistra: nessuna preselezione; a destra: selezione WP90.

Da come è evidente nella figura 3.10, il p_t dell'Higgs è una caratteristica di considerevole importanza insieme al p_t dei due fotoni.
Per cercare di aggirare questo problema è stato ricavato un nuovo campione aggregando tutti gli eventi provenienti dalle quattro differenti simulazioni prese in esame, in modo da avere una distribuzione del p_t approssimativamente più uniforme per il segnale, come mostrato in figura 3.11, dove si osservano dei picchi in corrispondenza a valori circa uguali alla metà della massa del T da cui decade il bosone di Higgs. Dai risultati riportati nella tabella 3.6, la configurazione *M tot* mostra un comportamento più stabile per il segnale BSM, anche se presenta ancora dei *bias*, seppur ridotti, rispetto al caso precedente, nel caso SM.

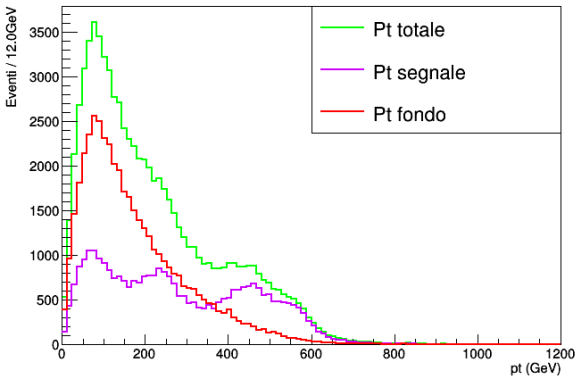


Figura 3.11: Distribuzione del p_t della configurazione *M tot*.

Per la configurazione *M tot train - 1000 test* è stata rappresentata in figura 3.12 la distribuzione di massa dell'Higgs in corrispondenza di punti di lavoro a 0.1%, a 1% e a 10%, che corrispondono rispettivamente ad un TPR al 3.4%, 28.8% e 92.7%. Infatti quest'ultimo appare come un picco ben definito su circa $125\ GeV$.

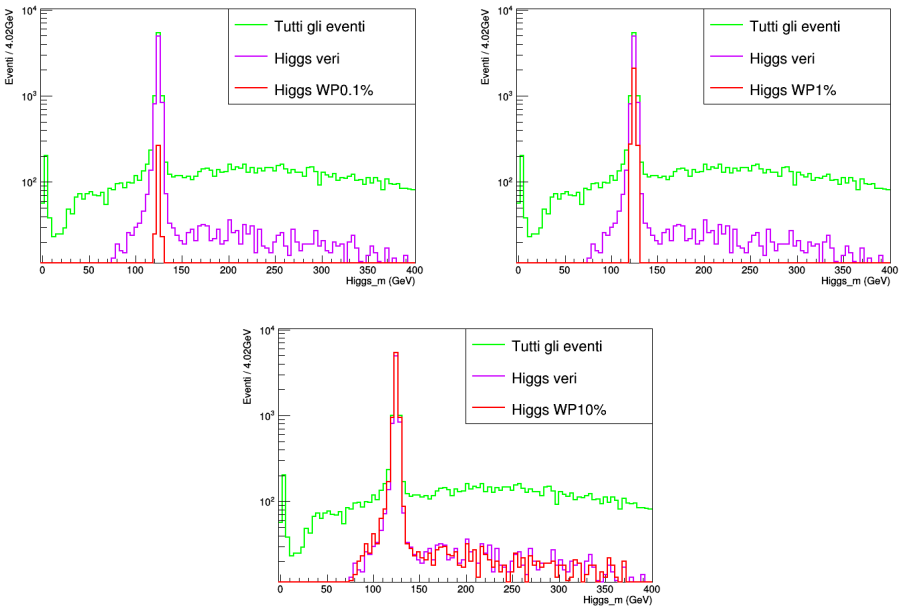


Figura 3.12: Massa del bosone di Higgs in fase di allenamento in corrispondenza di punti di lavoro a 0.1% (in alto a sinistra), a 1% (in alto a destra) e a 10% (in basso al centro) per la configurazione *M tot train - 1000 test*.

train - test	no sel.	WP90	WP80
M tot - M tot	95.6	88.7	87.9
M tot - 1200	96.2	87.7	86.9
M tot - 1000	95.9	87.1	87.1
M tot - 600	95.0	87.0	86.7
M tot - SM	90.8	84.9	84.6

Tabella 3.6: AUC in percentuale della configurazione *split* 70% train / 30% test e delle configurazioni miste dell'allenamento fatto sul campione M tot.

3.4.1 Risultati dell'allenamento: singolo fotone

La parte finale del lavoro verte sull'indagine dei fotoni del decadimento $H \to \gamma\gamma$ che appaiono sovrapposti a causa dell'elevato impulso del bosone di Higgs.
Talvolta è possibile che l'algoritmo *Particle Flow*, anziché identificare i due distinti fotoni, ne identifichi uno solo, il quale porta con sé tutte le informazioni sull'Higgs.
Il fotone in questione è da ricercare negli eventi in cui non viene identificato alcun candidato tramite l'algoritmo che ricostruisce i due fotoni trattato nel paragrafo 3.3.2. Di ogni fotone presente in tali eventi, è stato ricostruito il 4-vettore di Lorentz e il $\triangle R$ tra il fotone e il bosone di Higgs proveniente direttamente dalla simulazione prima del decadimento (*Higgs vero*). Per creare la *truth* da dare come input ad *xgboost*, nel segnale sono stati considerati i fotoni con una distanza $\triangle R < 0.4$ dall'Higgs vero, il restante è stato etichettato come fondo. Le catteristiche di questa nuova popolazione di candidati Higgs derivanti da un unico fotone sono: η, Φ, p_t, la massa, *electronVeto*, *pixelSeed*, *isScEtaEB*, *isScEtaEE*, *sieie*, *R9*, *pfRelIso03_all* e *pfRelIso03_chg*. Queste ultime variabili sono discusse in modo approfondito in [11].

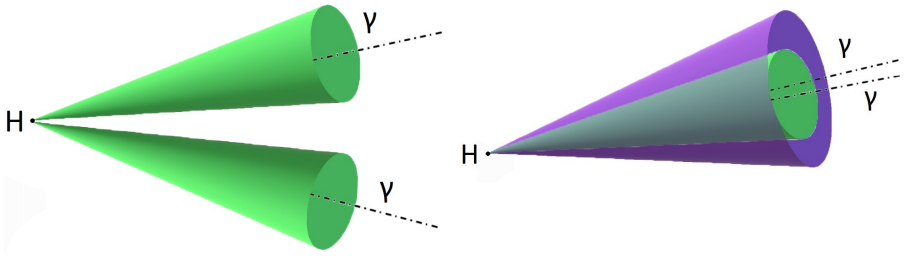


Figura 3.13: A sinistra: Il bosone di Higgs decade in due fotoni che sono ben distinguibili dal sistema di rilevazione; a destra: i fotoni prodotti dal decadimento dell'Higgs sono quasi sovrapposti e ricostruiti erroneamente in fatjet.

Inoltre, sono state ricercate le tracce di questi fotoni molto energetici indagando le caratteristiche dei *fatjet*, getti adronici che si distinguono dai jet poiché occupano una regione di spazio più vasta, rappresentabile come un cono come in figura 3.13. È possibile che i due γ ricostruiti si trovino all'interno della regione occupata dal fatjet in corrispondenza del fotone ricostruito.
Sono stati scartati dunque tutti i fatjet con $p_t < 50GeV$ e sono stati considerati, per il segnale, solo quelli con $\triangle R < 0.4$ dal fotone, il resto è stato considerato fondo. Dunque sono state aggiunte ulteriori due caratteristiche riferite ai fatjet così associati:

- m_softdrop: massa del fatjet corretta da un algoritmo chiamato PUPPI per la rimozione degli eventi *pile-up* [15];
- matched: variabile che contiene i valori di m_softdrop solo quando questa ha un valore maggiore di 100 GeV, il resto ha valore pari a -1.

Analogamente al caso precedente, il classificatore è stato allenato sia utilizzando *mvaID WP90* e *mvaID WP80* come caratteristiche, sia effettuando una selezione sui dati prima dell'allenamento con tali ID.
I risultati dell'allenamento mostrano una classificazione non ottimale dei dati di input rispetto a quelli ottenuti in 3.3.2, poiché, vista la natura complessa del segnale, la selezione descritta porta ad una contaminazione da eventi di segnale quando la categoria è quella di fondo.
Per risolvere il problema, il fondo è stato arricchito da due simulazioni dove non è presente il bosone di Higgs, in modo da essere sicuri che non ci fossero contaminazioni di segnale nel fondo. Tali simulazioni fanno riferimento ai processi rappresentati in figura 3.14.

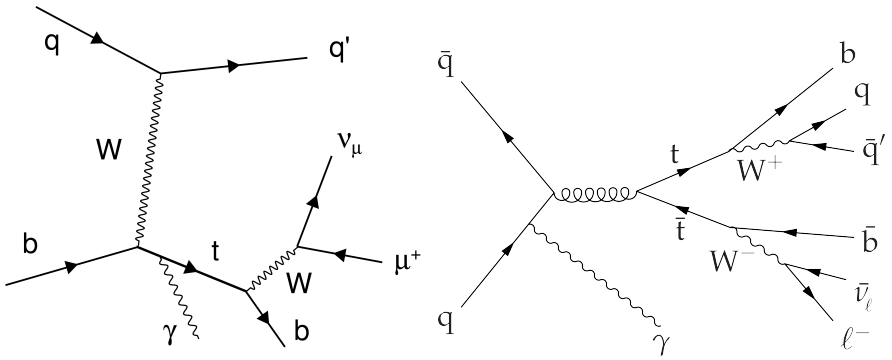


Figura 3.14: A sinistra: processo $t + \gamma$; a destra: processo $t\bar{t} + \gamma$.

I risultati di questo allenamento mostrano un'efficienza migliore del classificatore sia per le configurazioni *split*, sia per quelle miste. A titolo d'esempio sono

stati riportati i risultati della configurazione *SM train - 1200 test* nella figura 3.15 con le relative ROC 3.16. È evidente che ci sia un ottimo discernimento tra fondo e segnale, e le AUC suggeriscono che il miglior allenamento sia quello senza preselezione; anche in questo caso il classificatore ha prestazioni peggiori allenandosi su un set di dati composti dai candidati "migliori" sopravvissuti ai filtri WP.

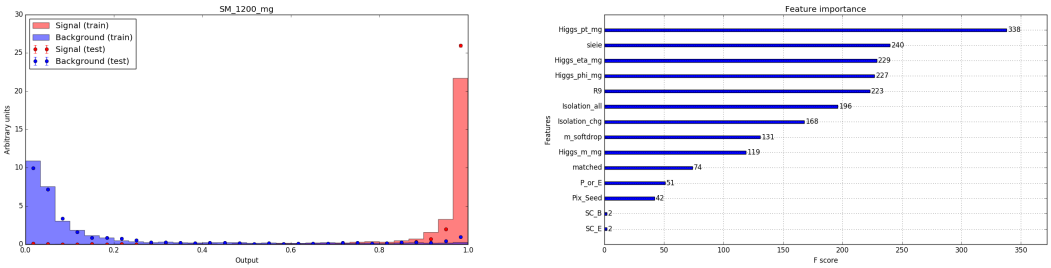


Figura 3.15: Risultati riferiti alla configurazione *SM train - 1200 test*. A sinistra: segnale e fondo; a destra: F-score delle variabili.

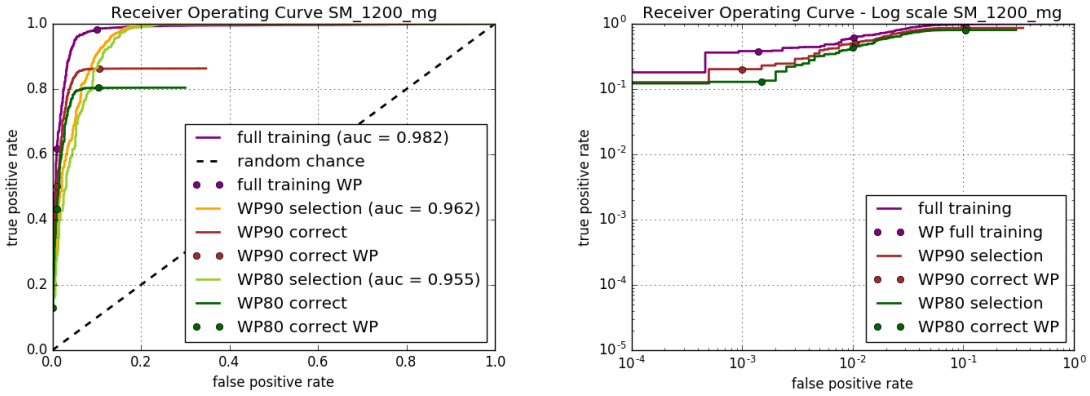


Figura 3.16: A sinistra: ROC delle configurazioni *SM train - 1200 test*; a destra: ROC in scala logaritmica.

Come vediamo dal grafico che riporta gli F-score delle caratteristiche più informative, il p_t dell'Higgs risulta essere la caratteristica più importante per il classificatore.
Per indagare il *bias* dovuto all'impulso trasverso sono stati effettuati due ulteriori allenamenti richiedendo che i candidati Higgs abbiano rispettivamente $p_t < 100\ GeV$ e $p_t > 100\ GeV$ per allenamento.

	$p_t < 100$			$p_t > 100$		
train - test	no sel.	WP90	WP80	no sel.	WP90	WP80
1200 - 1200	94.5	98.8	98.9	97.4	96.6	96.0
1000 - 1000	94.8	94.4	98.1	96.6	93.7	93.2
600 - 600	90.8	93.8	95.1	91.8	80.9	84.8
SM - SM	94.9	91.1	92.4	87.4	81.9	78.0
1200 - 1000	87.9	74.3	73.0	95.6	92.6	92.6
1200 - 600	81.8	82.8	84.2	84.3	72.7	72.4
1200 - SM	82.5	80.6	82.2	84.0	71.2	70.9
1000 - 1200	91.1	85.4	86.7	87.2	77.5	77.9
1000 - 600	81.1	85.4	86.1	87.1	77.5	77.9
1000 - SM	82.1	83.9	83.2	83.9	73.0	73.0
600 - 1200	78.9	81.5	83.8	80.5	66.9	65.7
600 - 1000	86.2	86.9	87.3	85.2	72.7	70.8
600 - SM	86.2	86.9	87.3	84.3	71.0	69.7
SM - 1200	78.9	84.6	86.5	89.2	78.2	79.4
SM - 1000	74.5	73.4	72.3	84.9	73.8	71.3
SM - 600 test	78.5	90.5	89.9	85.0	71.0	71.1

Tabella 3.7: AUC in percentuale per gli allenamenti con richiesta $p_t < 100\ GeV$ e $p_t > 100\ GeV$.

Per l'allenamento sul set di dati con $p_t > 100$ le AUC evidenziano prestazioni tendenzialmente migliori rispetto a quello con $p_t < 100$, salvo nel caso in cui si alleni e testi su configurazioni i cui Higgs candidati hanno un impulso trasverso più basso (SM e massa del T = 600). È possibile che selezionando soltanto i candidati con $p_t < 100$, la quantità di Higgs ben ricostruiti presenti nel segnale BSM si riduca al punto da inficiare sulle prestazioni; da ciò ne consegue che nella maggior parte dei casi la selezione WP80 o WP90 sembra che aiuti a costruire un segnale migliore che porta a valori di AUC più alti.
Al contrario, nel caso $p_t > 100$, essendo il segnale BSM ricostruito meglio, le preselezioni WP tenderanno a peggiorare le prestazioni del classificatore, facendogli perdere generalità.
Si può concludere dunque che gli allenamenti privi di selezione sono in ogni caso i più stabili. Anche per configurazioni *train - test* per cui c'è una grande differenza nel p_t dei candidati, il classificatore ha risultati buoni che non mostrano un *bias* eccessivo. A dimostrazione di ciò, sono stati riportati in figura 3.17 la distribuzione della *m_softdrop* a punti di lavoro fissati a 10%, a 1% e a 0.1%, i quali corrispondono rispettivamente ad un'efficienza del 12%, 84% e 90%.

Dal confronto con le distribuzioni di segnale è possibile notare che i picchi

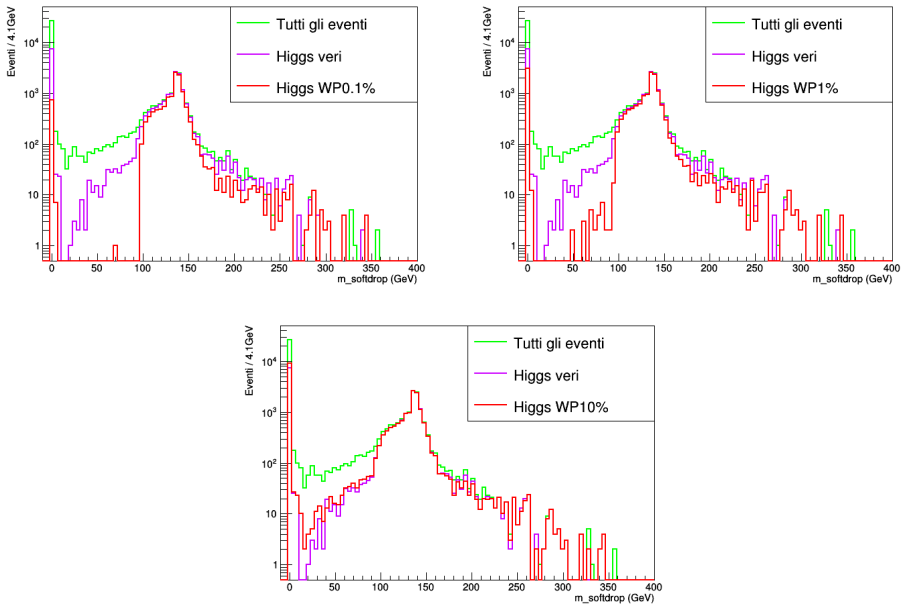


Figura 3.17: *m_softdrop* in fase di allenamento in corrispondenza di punti di lavoro a 0.1% (in alto a sinistra), a 1% (in alto a destra) e a 10% (in basso al centro) per la configurazione *SM train - 1200 test*.

intorno alla massa del bosone di Higgs risultano ben visibili e una buona parte dei fatjet a massa più bassa e appartenenti al fondo vengano rigettati dal classificatore.

Conclusioni

Nonostante il Modello Standard (MS) sia la teoria che oggi descrive al meglio il mondo subatomico, lascia irrisolti alcuni interrogativi a cui si cerca di rispondere attraverso la formalizzazione di modelli chiamati *Beyond Standard Model* (BSM). L'esperimento *Compact Muon Solenoid* (CMS) situato presso LHC al CERN ha come obiettivo l'indagine dei limiti del MS e la ricerca di nuova fisica teorizzata in tali modelli. Sia LHC che CMS sono progetti in continuo aggiornamento ed evoluzione tecnologica e la quantità di dati di collisione continuerà ad aumentare nel corso dei prossimi anni. In virtù della grande mole di dati che acquisisce CMS, deve fronteggiare il problema della sua elaborazione nel minor tempo e nel miglior modo possibile. A tal scopo, è sempre più frequente l'utilizzo di tecniche di *Machine Learning* (ML) per la ricostruzione delle particelle in esperimenti come CMS. Mediante tali tecniche, il presente lavoro di tesi presenta uno studio sulla ricostruzione del bosone di Higgs attraverso il canale di decadimento $H \to \gamma\gamma$.
È stato scelto un algoritmo supervisionato, dunque la prima fase del lavoro si incentra sulla ricostruzione dei dati da dare in input al ML, i quali sono stati generati partendo da campioni di dati simulati di CMS, corrispondenti alle condizioni di presa dati del Run II di LHC. In particolare, sono stati utilizzati campioni simulati di fenomeni BSM e SM; per i primi sono state indagate simulazioni del decadimento dell'ipotetico fermione T, vector-like quark il cui canale di decadimento in esame è $T \to Ht$, effettuando diverse ipotesi per la sua massa.
Tale ricostruzione avviene mediante delle richieste sulle caratteristiche dei bosoni di Higgs cercati; i candidati che superano tali richieste sono stati etichettati infine con una variabile booleana in modo da distinguere il segnale dal fondo e allenare l'algoritmo supervisionato.
Inoltre è stata effettuata la ricerca dei fotoni di tale decadimento che che appaiono sovrapposti a causa dell'elevato impulso del bosone di Higgs per estendere le casistiche contemplate dall'algoritmo di ricostruzione.
La seconda fase del lavoro riguarda l'addestramento dell'algoritmo, specificatamente è stato utilizzato un *Boosted Decision Tree* (BDT): un classificatore che, attraverso un processo iterativo, per ogni iterazione classifica le istanze di addestramento precedentemente mal modellate.

Per analizzare le prestazioni di tale algoritmo sono stati utilizzati dei modelli statistici basati sul *test d'ipotesi*. All'algoritmo sono state sottoposte varie configurazioni di dati, con e senza preselezioni fatte attraverso delle variabili risultanti da precedenti analisi multivariate, al fine di confrontare le prestazioni dei diversi allenamenti. È stato visto che il classificatore tende ad avere prestazioni migliori su dati non preselezionati là dove il segnale e il fondo sono ricostruiti al meglio. Da un confronto dei dati emerge un *bias* nell'allenamento in quanto gli Higgs analizzati, poiché provengono da simulazioni MS e BSM di differenti fenomeni fisici, presentano caratteristiche cinematiche tra loro differenti; in particolar modo ciò è da ricondursi all'impulso trasverso p_t dei bosoni candidati.
In ultima istanza possiamo affermare che i risultati sono molto promettenti e rappresentano un buon punto di partenza per l'ottimizzazione delle attuali tecniche di rivelazione del bosone di Higgs per questo canale di decadimento. Per migliorare ulteriormente tale analisi si potrebbe indagare il modo migliore per decorrelare la variabile p_t, in modo da rendere l'allenamento del BDT indipendente dal campione utilizzato.

Bibliografia

[1] Wikipedia. Il principio di simmetria nel modello standard. https://it.wikipedia.org/wiki/Modello_standard.

[2] P. W. HIGGS. Broken symmetries and the masses of gauge bosons. *JPhysical Review Letters, 13*, 1964.

[3] Aguilar-Saavedra, J. A. Handbook of vector-like quarks: Mixing and single production. *Phys. Rev. D 88, 094010*, 2013.

[4] Institute oh physics publishing and SISSA. LHC Machine. *JINST 3 S08001*, 2008.

[5] Journal of Instrumentation an IOP and SISSA journal. The cern large hadron collider: Accelerator and experiments. https://jinst.sissa.it/LHC/.

[6] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST 3 S08004*, 2008.

[7] CMS Collaboration. The CMS trigger system. *JINST 12 (2017) P01020*, 2015.

[8] CERN. Particle-flow reconstruction and global event description with the CMS detector. *Journal of Instrumentation*, 12, 2017.

[9] Project Management Committee (PMC). Introduction to boosted trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html.

[10] Luca Lista. *Statistical Methods For Data Analysis*. Springer, 2017.

[11] CMS Collaboration. Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions. *Journal of Instrumentation*, 10, 2015.

[12] CERN. About root. https://root.cern/about/.

[13] Project Management Committee (PMC). Python api reference of xgboost. https://xgboost.readthedocs.io/en/latest/python/python_api.html.

[14] NumFOCUS sponsored project. pandas - python data analysis library. https://pandas.pydata.org/.

[15] CERN for the benefit of the CMS collaboration. Pileup mitigation at CMS in 13 TeV data. *Journal of Instrumentation*, 2020.