Alma Mater Studiorum - Università di Bologna

Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Indagine sul consumo energetico e sul relativo impatto ambientale degli algoritmi di intelligenza artificiale

Tesi di laurea

Presentata da:
Lisa Stella Rizqallah

Relatore:
Prof. Matteo Franchini

Anno Accademico 2024-2025

A Giacomo Arlotti,
toccava a me guidarti in questa lettura.

Abstract

Lo sviluppo dell'intelligienza artificiale cresce sempre più rapidamente e così anche il consumo energetico delle infastrutture computazionali. Questo lavoro analizza i fondamenti del calcolo digitale e valuta l'impatto energetico degli algoritmi di AI. Vengono esaminati scenari applicativi scientifici ed impieghi quotidiani dell'AI, evidenziando l'impronta energetica e le implicazioni economiche e sociali. Infine, il lavoro propone strategie per ridurre il consumo energetico. L'obiettivo è fornire una visione integrata del rapporto tra AI, potenza computazionale e sostenibilità, sottolineando sfide e possibili soluzioni per uno sviluppo tecnologico più efficiente.

Indice

Introduzione

Il rapido sviluppo delle intelligenze artificiali (IA) ha determinato una crescita significativa della capacità di elaborazione e gestione dell'informazione. L'IA si sta progressivamente affermando come uno degli strumenti più rilevanti dell'era contemporanea, trovando applicazione in un'ampia varietà di ambiti che spaziano dalla ricerca scientifica all'industria, dai servizi digitali alla vita quotidiana. Grazie alla capacità di analizzare grandi quantità di dati e di individuare correlazioni complesse, i sistemi di intelligenza artificiale consentono oggi di affrontare problemi che risultavano precedentemente difficilmente trattabili con metodi computazionali tradizionali.
Il progresso degli algoritmi di apprendimento automatico, ed in particolare delle tecniche di deep learning, è stato reso possibile da un insieme di diversi fattori tecnologici: l'aumento esponenziale della disponibilità di dati digitali, lo sviluppo di architetture hardware sempre più performanti e la crescente diffusione di infrastrutture di calcolo ad alte prestazioni. Tuttavia, tali progressi sono accompagnati da un incremento significativo delle risorse computazionali richieste per l'addestramento e l'utilizzo dei modelli di intelligenza artificiale. Di conseguenza, il consumo energetico associato a queste tecnologie rappresenta tuttora uno degli aspetti più critici e discussi nell'ambito dello sviluppo sostenibile dell'ecosistema digitale.
Infatti, il funzionamento continuo dei data center e l'addestramento di modelli di grandi dimensioni, comportano un fabbisogno energetico considerevole con implicazioni non trascurabili sia dal punto di vista economico sia dal punto di vista ambientale. L'energia necessaria per alimentare le infrastrutture di calcolo e per garantire il raffreddamento dei sistemi hardware contribuisce massivamente alle emissioni globali di gas serra, soprattutto quando l'approvvigionamento energetico dipende in larga misura da fonti fossili. Dunque, l'analisi dell'impatto energetico dell'intelligenza artificiale assume un ruolo centrale non solo nella valutazione dell'efficienza tecnologica dei sistemi, ma anche nella definizione di strategie di sviluppo sostenibile.
Inoltre, l'IA non rappresenta soltanto una fonte di consumo energetico, ma può costituire anche uno strumento fondamentale per l'ottimizzazione dei sistemi energetici stessi: sono già impiegate tecniche di apprendimento automatico nella gestione delle reti elettriche, nell'ottimizzazione dei processi industriali e nella riduzione degli sprechi energetici. Questa duplice natura dell'intelligenza artificiale, al tempo stesso tecnologia energivora e potenziale strumento per una maggiore efficienza, rende necessario un approccio equilibrato che tenga conto sia dei benefici sia dei costi ambientali associati alla sua diffusione.

L'obiettivo della presente tesi è quindi indagare il rapporto tra sviluppo del-

l'IA e consumo energetico esaminando i principali fattori che determinano l'impatto energetico e le possibili strategie per ridurne l'impronta ambientale. È importante sottolineare che la presente tesi non intende presentare un'analisi esaustiva di tutte le applicazioni dell'intelligenza artificiale né di quantificare con precisione assoluta i consumi energetici di ogni sistema. Si tratta piuttosto di uno studio divulgativo/descrittivo che si propone di fornire una panoramica delle architetture di calcolo utilizzate nei sistemi di intelligenza artificiale, con una sezione quantitativa, in cui si presentano dati di consumo e stime energetiche per illustrare l'impatto reale dei sistemi associato alle diverse fasi di funzionamento degli algoritmi; e infine con una parte prospettica, che discute possibili soluzioni tecnologiche e strategie di sostenibilità, includendo scenari emergenti come il quantum computing e le architetture neuromorfiche, orientate al miglioramento dell'efficienza energetica.
Nel corso della trattazione verranno inoltre illustrati diversi ambiti applicativi, con particolare attenzione alle applicazioni tecnico-scientifiche dell'intelligenza artificiale che rappresenta uno strumento di analisi estremamente potente, capace di accelerare l'elaborazione dei dati sperimentali e di contribuire allo sviluppo di nuovi modelli interpretativi dei fenomeni fisici.
La struttura della tesi è organizzata in sei capitoli. Nel primo, vengono introdotti i fondamenti del calcolo digitale e le principali architetture hardware impiegate nei sistemi di intelligenza artificiale, con particolare attenzione al loro ruolo nel determinare il consumo energetico complessivo. Il secondo capitolo analizza in maniera più approfondita l'impatto energetico degli algoritmi di IA, distinguendo tra le diverse fasi di funzionamento dei modelli e discutendo le principali metriche utilizzate per la valutazione dei consumi. Nel terzo, vengono presentate alcune applicazioni in ambito scientifico sottolineando il contributo dell'intelligenza artificiale alla ricerca. Il quarto capitolo invece, si concentra sull'utilizzo dell'IA nella vita quotidiana e sulle implicazioni sociali ed economiche associate alla diffusione collettiva dei servizi digitali basati su algoritmi di apprendimento automatico. Nel quinto capitolo vengono infine discusse diverse strategie per la riduzione dell'impatto energetico dell'intelligenza artificiale, mentre il sesto capitolo, invece, introduce l'emergente Quantum Computing, analizzandone il potenziale in termini di efficienza computazionale ed energetica.

Attraverso questa indagine, la tesi intende analizzare, diffondere e contribuire alla comprensione delle sfide energetiche associate allo sviluppo dell'intelligenza artificiale e delle possibili direzioni di ricerca orientate alla realizzazione di tecnologie digitali più sostenibili.

Capitolo 1

Energia e computazione

Non esiste una definizione univoca ed universalmente accettata di intelligenza artificiale (IA): in termini generali, può essere descritta come l'insieme di metodologie e modelli computazionali finalizzati alla realizzazione di sistemi in grado di apprendere ed eseguire compiti che tradizionalmente richiedono capacità cognitive umane. Però, a differenza degli approcci computazionali classici, basati esclusivamente su algoritmi deterministici e istruzioni esplicitamente programmate, l'IA moderna è fondata su tecniche di apprendimento dai dati, attraverso le quali i sistemi sono in grado di individuare strutture e correlazioni, formulare previsioni ed infine effettuare decisioni operative. Le prestazioni dei sistemi di IA migliorano progressivamente mediante processi di addestramento ed ottimizzazione.

Nonostante l'intelligenza artificiale venga spesso descritta in termini di capacità funzionali e algoritmiche, la sua realizzazione concreta è legata ai principi fondamentali del calcolo automatico e alle architetture fisiche che lo rendono possibile. I processi di apprendimento, inferenza e ottimizzazione, pur nella loro complessità, sono implementati attraverso operazioni elementari di elaborazione dell'informazione. Infatti, comprendere il funzionamento dell'IA richiede un'analisi dei meccanismi di rappresentazione, codifica e manipolazione dei dati alla base del calcolo digitale, che costituiscono il substrato logico e materiale su cui tali algoritmi operano.

1.1 Fondamenti del calcolo: dal sistema decimale alla logica binaria

L'evoluzione dell'architettura del calcolo su scala globale ha imposto un'importante divergenza dai metodi computazionali tipici della matematica decimale. Mentre il sistema numerico decimale si configura come una notazione posizionale in base 10, che impiega appunto dieci cifre 0,1,...,9 e segue una logica di conteggio ciclica, il calcolo automatico moderno adotta esclusivamente il sistema binario basato su due sole cifre. Infatti, il sistema binario è una notazione posizionale in base 2 dove il conteggio è ristretto a due soli stati discreti, denotati come 0 e 1 ed utilizza combinazioni delle due cifre disponibili: il numero successivo a 1 è 10 (analogamente a quanto avviene nel sistema decimale dopo il 9). È possibile stabilire una corri-

spondenza univoca tra i numeri decimali e quelli binari: 0 e 1 rimangono invariati, mentre il numero decimale 2 viene rappresentato come 10 in base 2 (10_2), seguito da 11 per il 3, 100 per il 4, 101 per il 5, e così via: la progressione numerica avvenga per saturazione di bit. La scomposizione di un numero intero N in base 2 segue infatti la legge di potenza:

$$N = \sum_{i=0}^{n} a_i 2^i \quad a_i \in \{0, 1\} \tag{1.1}$$

Il valore intrinseco del sistema binario risiede nella sua immediata traducibilità in termini di algebra di Boole. Le cifre 0 e 1 vengono interpretate come stati logici rispettivamente di Falso e Vero, permettendo la costruzione di funzioni logiche complesse attraverso operatori fondamentali (AND, OR, NOT). Questa astrazione è implementata mediante la discretizzazione di segnali elettrici.

Un bit (binary digit), unità fondamentale dell'informazione, viene associato a specifici livelli di potenziale elettrico all'interno di circuiti a semiconduttore:

- Stato 0: Assenza di tensione o livello di soglia inferiore (V_{low}).
- Stato 1: Presenza di tensione o livello di soglia superiore (V_{high}).

Questa corrispondenza tra logica matematica e stati quantizzati della materia permette ai dispositivi elettronici di elaborare, codificare e trasmettere dati complessi, inclusi i caratteri testuali, mediati da standard di codifica quali ASCII o UTF-8. Il bit ammette numerosi multipli, l'unità di misura derivata di maggiore utilizzo è il byte, definito come un ottetto di bit (1 byte=8 bit). In realtà, data l'espansione esponenziale dell'ecosistema digitale, è necessario ricorrere a multipli di ordine superiore. Tra questi, lo zettabyte (ZB) rappresenta una grandezza con una duplice definizione per via di differenti standard di riferimento:

- Standard SI (Sistema Decimale): 1 ZB = 10^{21} byte.
- Standard IEC (Binario): 1 ZB = 2^{70} byte.

Un'analisi condotta pochi anni fa dalla società di ricerche di mercato IDC ([1]) ha fornito una stima indicativa dell'ampiezza dell'universo digitale che evidenzia una crescita asintotica della produzione globale di dati. Nel 2013, la quantità totale di informazione digitale prodotta è stata stimata in circa 4,3 zettabyte, con una previsione, confermata dalle analisi successive, di un incremento di un ordine di grandezza fino a 44 zettabyte entro il 2020. Per cercare di contestualizzare e rendere più intuitivo l'ordine di grandezza di questi volumi di informazione, è possibile considerare alcuni esempi concreti: se tali dati fossero memorizzati su supporti ottici come i DVD, la loro disposizione in sequenza coprirebbe due volte la distanza media tra la Terra e la Luna ($\approx 7.7\times105$ km). Analogamente, utilizzando smartphone come unità di archiviazione, si potrebbe realizzare una catena di dispositivi sufficiente ad avvolgere l'intera circonferenza terrestre per oltre 120 volte. [2]

1.2 Architetture di calcolo: unità ed evoluzione. Confronto tra CPU, GPU, TPU ed unità specializzate

Nell'ambito dell'intelligenza artificiale, le diverse architetture di calcolo rivestono un ruolo particolarmente rilevante per quanto riguarda l'ottimizzazione delle prestazioni e l'efficienza dei modelli. Le diverse unità di elaborazione sono infatti progettate per rispondere ad esigenze computazionali differenti.

- CPU (Central Processing Unit): componente fondamentale per il calcolo sequenziale e le operazioni. È architettata con un numero limitato di core ma con un'alta frequenza di clock (risposta immediata per rapida esecuzione delle istruzioni) e con una cache veloce (rapido accesso ai dati maggiormente utilizzati). È in grado di gestire efficacemente documenti e navigazione nel web, simultanei processi (multitasking) e calcoli complessi con algoritmi avanzati.

- GPU (Graphics Processing Unit): progettata per gestire compiti grafici ed elaborazione dati in parallelo (elabora enormi quantità di dati simultaneamente). Composta da migliaia di core ma con frequenze operanti inferiori a quelle della CPU. In particolare, crea animazioni fluide ed applica rapidamente effetti speciali nell'editing video, possiede grande spazio per gestire dati grafici complessi.

Accanto a queste architetture consolidate, stanno anche emergendo nuove unità di elaborazione specializzate, le TPU, sviluppate specificamente per operazioni su tensori, le più recenti NPU (Neural Processing Unit) e le LPU (Language Processing Unit). Tali soluzioni sono progettate con il fine di migliorare l'efficienza energetica e ridurre la latenza in compiti specifici, consentendo così l'implementazione di applicazioni sempre più complesse e sofisticate, anche in contesti con risorse computazionali limitate.

- TPU: sviluppate da Google, accelerano il lavoro legato all'apprendimento automatico e all'intelligenza artificiale. Ottimizzate per i calcoli tensoriali, base del deep e machine learning, e performanti nell'esecuzione di inferenza delle reti neurali profonde (processo che un modello di apprendimento automatico addestrato utilizza per trarre conclusioni da dati nuovi) con maggiore efficienza computazionali e minore consumo energetico.

- NPU: efficiente esecuzione delle operazioni neuronali con velocità di esecuzione molto elevata e ridotto consumo energetico. Accelerano le operazioni del deep learning come convoluzione e attivazione offrendo prestazioni superiori rispetto alle altre unità di elaborazione, in particolare in termini di versatilità e flessibilità rispetto alle TPU.

- LPU: progettate per analizzare il linguaggio umano (elaborazione del testo e comprensione dell'interazione tra essere umani e computer). Raggiungono prestazioni avanzate anche nel riconoscimento vocale, nella traduzione automatica ma anche nella flessibilità dei diversi scenari e nell'adattamento alla

vasta gamma di dispositivi a cui vengono integrate, rendendole superiori ad altre unità di elaborazione.

L'analisi comparativa permette dunque di valutare i punti di forza ed i limiti applicativi, fornendo un quadro utile per la scelta dell'architettura più adeguata in funzione del tipo di algoritmo di intelligenza artificiale considerato. [3]
In Tabella 1.1, sono inoltre riportati parametri come il numero di core, la frequenza tipica, il consumo medio ed il workload principale per ogni tipologia di unità di elaborazione. [4], [5], [6]

Unità	**Core**	**Frequenza**	**Consumo**	**Workload principale**
CPU	4–64	2–5 GHz	65–250 W	Calcolo sequenziale, multitasking
GPU	1.000–18.000	1–2 GHz	150–700 W	Grafica, deep learning, calcolo parallelo
TPU	∼65.000 MAC*	∼0.7 GHz	75–200 W	Addestramento e inferenza ML
NPU	decine–centinaia	∼0.5–1 GHz	0.5–10 W	Inferenza AI su dispositivi edge
LPU	arch. specializzata	∼1 GHz	50–200 W	Elaborazione linguaggio naturale

Tabella 1.1: Confronto tra diverse unità di elaborazione utilizzate nei sistemi di intelligenza artificiale.

*MAC (*Multiply–Accumulate unit*) è un'unità hardware che esegue in un singolo ciclo l'operazione $a = a + (b \times c)$, fondamentale per moltiplicazioni matrice-matrice, moltiplicazioni matrice-vettore e convoluzioni nelle reti neurali.

1.3 Consumo energetico

1.3.1 Impatto ambientale

Poiché i sistemi di intelligenza artificiale operano attraverso l'elaborazione di grandi quantità di informazione mediante architetture di calcolo digitali, il loro funzionamento è obbligatoriamente associato ad un consumo di energia. La crescente complessità degli algoritmi di IA e delle infrastrutture hardware su cui vengono eseguiti ha reso il fabbisogno energetico uno degli aspetti più dibattuti nello sviluppo e nello studio di tali tecnologie.
Il consumo energetico associato agli algoritmi di intelligenza artificiale non si traduce esclusivamente in un aumento dei costi operativi, ma comporta anche rilevanti conseguenze dal punto di vista ambientale. Una parte significativa dell'energia impiegata per alimentare i data center proviene da fonti non rinnovabili, contribuendo alle emissioni di anidride carbonica e all'intensificazione dei fenomeni legati al cambiamento climatico in atto. Infatti, l'ampio utilizzo dell'IA da parte degli utenti

comporta un importante aumento della domanda energetica, con implicazioni ambientali non trascurabili. Diventa quindi essenziale valutare e mitigare tali emissioni al fine di garantire uno sviluppo sostenibile delle tecnologie.
L'addestramento di modelli di IA avanzati e la gestione di grandi volumi di dati sempre maggiori richiedono risorse computazionali estremamente elevate. In particolare, i recenti progressi della ricerca hanno favorito la diffusione di algoritmi di deep learning caratterizzati da un'elevata complessità computazionale e di conseguenza da un intenso consumo di energia.
Secondo uno studio condotto dal Carbon Trust, organizzazione no-profit specializzata in sostenibilità ambientale, il funzionamento continuo di un data center di medie dimensioni per un anno può generare emissioni di CO_2 dell'ordine di diverse centinaia di tonnellate, principalmente a causa dell'elevata dipendenza da energia elettrica prodotta mediante fonti fossili. Inoltre, stime recenti riportano che entro il 2030 l'intelligenza artificiale potrebbe arrivare a rappresentare fino al 4% del consumo globale di energia elettrica. Attualmente, aziende come Google, riportano che una quota compresa tra il 10% e il 15% del loro fabbisogno energetico totale è attribuibile alle applicazioni di IA, corrispondente a circa 2,3 TWh annui (vedi grafico in Figura 1.1). Ulteriori analisi volte a quantificare l'impatto energetico dell'IA, indicano che il mantenimento degli attuali trend di sviluppo e adozione tecnologica potrebbe portare aziende come Nvidia a immettere sul mercato circa 1,5 milioni di unità server. Queste infrastrutture, se operate a pieno regime, richiederebbero un consumo energetico annuale stimato in almeno 85,4 TWh, un valore di gran lunga superiore al fabbisogno elettrico annuo di numerosi stati di piccole dimensioni. [7]

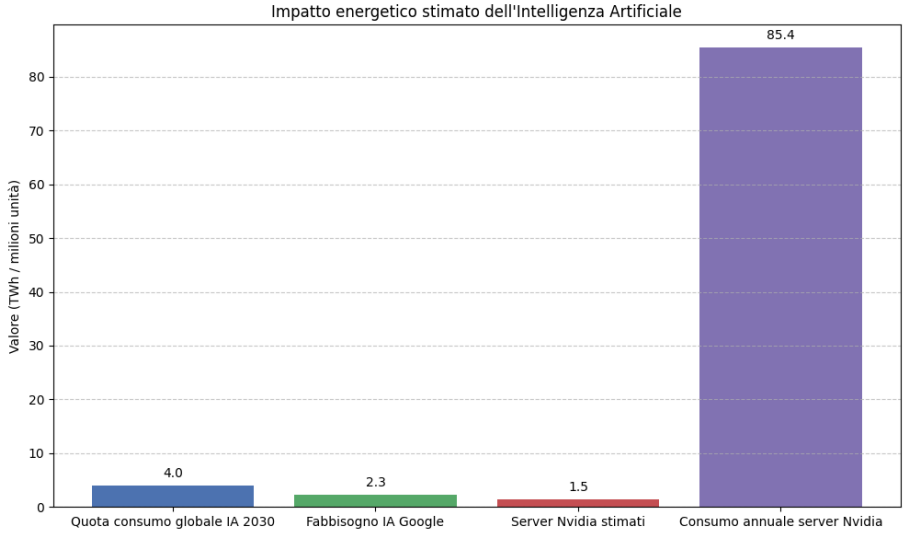

Figura 1.1: Grafico dell'impatto energetico dell'IA.

1.3.2 Stima emissioni di CO_2: unità di misura e calcolo

Per stimare le emissioni di anidride carbonica associate al consumo di energia elettrica è necessario introdurre un fattore di conversione espresso in termini di kg di CO_2 per kilowattora (kg CO_2/kWh). Questo fattore quantifica la quantità di anidride carbonica emessa in atmosfera per ogni unità di energia elettrica consumata.

Esistono differenti metodologie per il calcolo di questo parametro, tra le quali le più comunemente adottate sono l'approccio location-based e l'approccio market-based. A seconda del metodo utilizzato, il valore del fattore di emissione, espresso in kg CO_2e/kWh, può variare in modo anche significativo.

L'approccio location-based stima le emissioni di CO_2 per kWh sulla base del mix energetico medio della rete elettrica nazionale. Il fattore di emissione rappresenta il valore medio delle emissioni associate alla produzione di 1 kWh di energia elettrica, considerando l'insieme delle fonti primarie che contribuiscono alla generazione elettrica del Paese. Poiché l'energia immessa in rete proviene da fonti eterogenee quali gas naturale, carbone, impianti fotovoltaici, eolici e idroelettrici, e una volta distribuita non è possibile risalire all'origine specifica, questo metodo fornisce una stima aggregata e rappresentativa del reale prelievo dalla rete elettrica.
Il principale vantaggio dell'approccio location-based risiede nella sua capacità di riflettere fedelmente l'impatto ambientale dell'energia effettivamente consumata dalla rete.
Tuttavia, un limite di tale metodo è costituito dal fatto che la riduzione delle emissioni di CO_2 risulta possibile esclusivamente attraverso una diminuzione del consumo complessivo di energia elettrica, obiettivo non sempre facilmente conseguibile.
L'approccio market-based, invece, stima le emissioni associate a 1 kWh di energia elettrica in funzione delle caratteristiche del fornitore scelto e del relativo mix di generazione. Questo metodo consente di valorizzare l'impatto positivo della scelta di fornitori che offrono energia proveniente in misura maggiore da fonti rinnovabili, rispetto a quelli che si basano prevalentemente su combustibili fossili. Così facendo, l'approccio market-based mette in evidenza come decisioni di approvvigionamento energetico possano incidere direttamente sulle emissioni attribuite ai consumi elettrici.

Per il calcolo delle emissioni associate al consumo di energia elettrica si utilizzano specifici fattori di emissione. Alcuni di essi considerano esclusivamente le emissioni di anidride carbonica (CO_2), mentre altri includono l'insieme di gas quali metano (CH_4) e protossido di azoto (N_2O). In quest'ultimo caso, le emissioni vengono espresse in termini di CO_2 equivalente (CO_2e), una grandezza che consente di confrontare e aggregare l'effetto climatico dei diversi gas serra sulla base del loro potenziale di riscaldamento globale. [8]

Capitolo 2

Impatto energetico

2.1 Perchè l'intelligenza artificiale consuma energia

L'addestramento e l'utilizzo di modelli di intelligenza artificiale avanzati, come ChatGPT o Google Gemini, comportano un fabbisogno di risorse energetiche e idriche non trascurabile. In particolare, il funzionamento dei data center dedicati a tali applicazioni richiede importanti quantità di energia elettrica per sostenere le operazioni di calcolo e l'enorme consumo di acqua destinata ai sistemi di raffreddamento necessari a garantire l'efficienza di tutte le infrastrutture hardware.
Gli algoritmi di intelligenza artificiale, in particolare quelli basati su reti neurali profonde (deep learning), sfruttano in modo intensivo entrambe queste componenti, determinando un significativo impatto energetico complessivo. Con l'aumento progressivo delle capacità computazionali delle tecnologie di intelligenza artificiale e la crescita delle richieste da parte degli utenti, il fabbisogno complessivo di risorse energetiche e materiali risulta in costante incremento.

Un'analisi condotta dall'Öko-Institute per conto di Greenpeace ([9]) evidenzia come il consumo globale di acqua destinata ai sistemi di raffreddamento dei data center sia destinato a crescere passando da circa 175 miliardi di litri nel 2023 a 664 miliardi di litri entro il 2030. Tale valore corrisponde approssimativamente al consumo annuo di acqua potabile di una città con una popolazione pari a circa tre volte quella di Milano.
Inoltre, l'espansione delle infrastrutture di calcolo e delle capacità di intelligenza artificiale potrebbe determinare un incremento fino a cinque milioni di tonnellate di rifiuti elettronici aggiuntivi entro lo stesso orizzonte temporale, sollevando ulteriori criticità in termini di sostenibilità ambientale e gestione delle risorse.
Stime recenti, riportate anche dalla Banca Centrale Europea, indicano che la generazione di una singola risposta da parte di un sistema di intelligenza artificiale può comportare un consumo energetico fino a 10 volte superiore rispetto a quello associato ad una tradizionale ricerca effettuata tramite un motore come Google. Infatti, sempre secondo le valutazioni dell'Öko-Institute, i data center specificamente progettati per applicazioni di intelligenza artificiale, presentano un fabbisogno idrico pari a circa il doppio rispetto a quello dei data center convenzionali, a causa delle maggiori esigenze di dissipazione del calore generate dall'elevata densità di potenza

dei sistemi di calcolo impiegati.

Anche la fase di produzione dell'hardware destinato alle applicazioni di intelligenza artificiale comporta un impatto ambientale significativo.
Secondo un rapporto di Greenpeace (come mostrato in Tabella 2.1), nel periodo compreso tra il 2023 e il 2024 il consumo di energia elettrica e le emissioni di gas serra associate alla produzione globale di chip per l'intelligenza artificiale hanno registrato un aumento rispettivamente del 351% e del 357%. Tali incrementi riflettono l'espansione sempre maggiore della domanda di componenti elettronici ad alte prestazioni necessaria a sostenere lo sviluppo dei sistemi di IA.

	Incremento percentuale	**Località**
Consumo di energia elettrica	+351%	Globale
Emissioni di gas serra	+357%	Globale

Tabella 2.1: Impatto ambientale della produzione di chip per l'IA, Greenpeace (2023-2024)

In particolare, l'Asia orientale, che rappresenta uno dei principali poli mondiali per la produzione di semiconduttori, risulta fortemente esposta alle conseguenze ambientali di questa crescita: la produzione di chip richiede ingenti quantità di energia elettrica, soddisfatte prevalentemente attraverso fonti fossili. A titolo esemplificativo, queste coprono circa il 58,5% del mix elettrico in Corea del Sud, il 68,6% in Giappone e l'83,1% a Taiwan (vedi Tabella 2.2). La forte dipendenza da combustibili fossili accentua l' impatto climatico associato alla filiera produttiva dell' hardware per l'intelligenza artificiale, evidenziando la necessità di strategie di decarbonizzazione. [10]

Fabbisogno elettrico percentuale	**Località**
58,5%	Corea del Sud
68,6%	Giappone
83,1%	Taiwan

Tabella 2.2: Percentuale di energia elettrica utilizzata per la produzione dei chip proveniente da fonti fossili, Greenpeace (2023-2024)

2.2 Domanda energetica dell'AI

2.2.1 Data center

La maggior parte delle attività legate all' addestramento e all' implementazione dei modelli di intelligenza artificiale avviene all' interno dei già citati data center, elementi centrali nell'interazione tra infrastrutture digitali e sistema energetico. Un

data center è una struttura progettata per ospitare grandi quantità di apparecchiature informatiche, tra cui principalmente server, sistemi di archiviazione e dispositivi di rete insieme a numerosi sistemi ausiliari necessari per garantirne il funzionamento continuo e affidabile.
I data center sono relativamente recenti all'interno del sistema energetico globale. Attualmente, il consumo di elettricità attribuibile a è stimato in circa 415 terawattora (TWh), pari a circa l'1,5% del consumo elettrico mondiale nel 2024. Negli ultimi cinque anni tale valore è cresciuto con un tasso medio annuo di circa il 12%, riflettendo la rapida espansione dei servizi digitali e delle applicazioni basate sui dati.

La componente principale è costituita dai server, ovvero computer dedicati all' elaborazione e alla gestione dei dati eventualmente dotati di unità di elaborazione centrali (CPU) e di acceleratori specializzati, come le unità di elaborazione grafica (GPU), utilizzate soprattutto nei carichi di lavoro legati all'intelligenza artificiale e all'apprendimento automatico.
Nei data center moderni i server rappresentano mediamente circa il 60% del consumo totale di elettricità, anche se questo valore può variare a seconda della tipologia e della configurazione della struttura.
Accanto ai server sono presenti i sistemi di archiviazione, utilizzati per la memorizzazione dei dati e per le operazioni di backup che contribuiscono a circa il 5% del consumo energetico complessivo.
Un' altra importante componente è composta dalle apparecchiature di rete che includono dispositivi come switch, router e sistemi di bilanciamento del traffico: gli switch collegano tra loro i dispositivi nel data center, i router gestiscono il traffico di rete, mentre i load balancer ottimizzano le prestazioni distribuendo i carichi di lavoro. Nel loro insieme, anche queste infrastrutture rappresentano fino a circa il 5% della domanda di elettricità.
Inoltre, sono fondamentali i sistemi di raffreddamento e controllo ambientale necessari per mantenere la temperatura entro intervalli ottimali e garantire il corretto funzionamento delle apparecchiature informatiche. Il consumo energetico associato al raffreddamento varia a seconda dell'efficienza del data center: nelle strutture hyperscale più efficienti può rappresentare circa il 7% del consumo totale, mentre nei data center meno efficienti può superare il 30%.
Per garantire un elevato livello di affidabilità sono inoltre presenti sistemi di alimentazione di emergenza, come batterie di continuità (UPS) e generatori di backup. Questi dispositivi entrano in funzione solo in caso di interruzioni dell'alimentazione elettrica, ma sono fondamentali per assicurare la continuità operativa richiesta dalle infrastrutture digitali critiche.

La stima del consumo energetico dei data center è caratterizzata da un significativo e non trascurabile grado di incertezza: l'analisi della domanda di elettricità associata allo sviluppo dell'IA richiede un approccio basato su scenari che consentono di esplorare possibili traiettorie alternative e di valutare i tempi di evoluzione rilevanti per il settore energetico. Per tener conto di queste incertezze, vengono considerate ipotesi differenti riguardo ai miglioramenti di efficienza hardware e software, al ritmo di adozione delle tecnologie di intelligenza artificiale e ai possibili vincoli infrastrutturali del sistema energetico.

In uno scenario di riferimento (Base Case), le proiezioni indicano che il consumo globale di elettricità dei data center potrebbe raddoppiare entro il 2030, raggiungendo circa 945 TWh, pari a poco meno del 3% della domanda elettrica mondiale. Nel periodo compreso tra il 2024 e il 2030, ciò corrisponderebbe a un tasso di crescita medio annuo di circa il 15%, più di quattro volte superiore rispetto alla crescita della domanda di elettricità degli altri settori. Nonostante questo importante incremento, la quota complessiva dei data center sul consumo elettrico globale rimarrebbe comunque relativamente contenuta.
Una parte rilevante di questa crescita è attribuibile ai server accelerati, il cui consumo elettrico è previsto aumentare con un tasso medio annuo di circa il 30%, mentre per i server convenzionali la crescita risulterebbe più moderata, attorno al 9% annuo. I server accelerati contribuirebbero quindi a quasi la metà dell'aumento del consumo energetico dei data center, mentre i server tradizionali rappresenterebbero circa il 20% dell'incremento. Le altre apparecchiature informatiche (vedi Figura 2.1), contribuiscono con circa il 10% e le infrastrutture di supporto, tra cui i sistemi di raffreddamento e le altre componenti non direttamente legate al calcolo, con circa il 20%. [11]

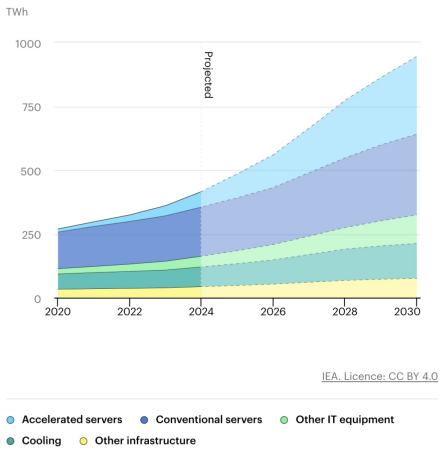

Figura 2.1: Consumo globale di elettricità dei data center suddiviso per tipo di apparecchiatura (Base Case), 2020-2030. [11]

La Figura 2.2 mostra invece l'aumento complessivo della domanda di elettricità coinvolge tutte le principali tipologie di data center, inclusi quelli aziendali (enterprise), le strutture di colocation e i provider di servizi ed i grandi data center hyperscale, progettati per supportare piattaforme cloud e servizi digitali su scala globale.

2.2.2 Potenza assorbita dagli algoritmi di AI

La potenza richiesta dagli algoritmi di intelligenza artificiale dipende da più fattori, tra cui la complessità architetturale del modello, la dimensione del dataset utilizzato e la frequenza delle operazioni di addestramento o inferenza. In particolare, la fase di addestramento necessita di una domanda energetica significativamente più elevata e

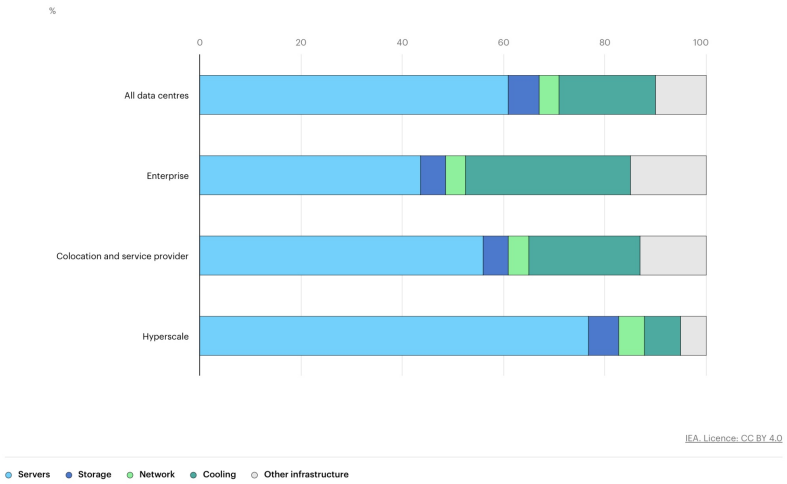

Figura 2.2: Ripartizione del consumo di elettricità nei data center per tipologia di apparecchiatura (2024). [11]

concentrata rispetto alle applicazioni tradizionali dei data center. Studi focalizzati su questa fase, tra cui l'analisi condotta da ricercatori di Google e dell'Università della California a Berkeley ([12]), indicano che il consumo energetico necessario per l'addestramento di alcuni modelli di intelligenza artificiale si colloca nell'ordine delle centinaia di megawattora. Tali dati risultano particolarmente rilevanti se si considera che essi si riferiscono esclusivamente alla fase di addestramento, alla quale devono poi essere sommati i consumi energetici associati alla fase di elaborazione di ogni singola richiesta.
Infatti, i modelli di intelligenza artificiale presentano un consumo energetico rilevante non solo durante la fase di addestramento (training), ma anche nel corso della loro operatività quotidiana, tecnicamente definita fase di inferenza (vedi paragrafo successivo). Di conseguenza, oltre alle emissioni di carbonio associate alla costruzione, all'addestramento e al perfezionamento dei modelli, è necessario considerare anche quelle generate durante il loro utilizzo operativo su larga scala.

Dati recenti indicano che la generazione di una singola immagine tramite un sistema di intelligenza artificiale generativa implica un consumo energetico paragonabile a quello richiesto per una ricarica completa di uno smartphone. Inoltre, se un operatore globale come Google dovesse sostituire totalmente il proprio servizio di ricerca tradizionale con un modello di ricerca basato su intelligenza artificiale, il consumo energetico complessivo associato a tale servizio potrebbe raggiungere valori dell'ordine di 29,3 TWh all'anno.
Come evidenzia un recente rapporto dell'Agenzia Internazionale per l'Energia, il consumo elettrico dei data center è infatti destinato a registrare una crescita significativa nel corso dei prossimi anni. Tale incremento è attribuibile principalmente all'aumento della domanda computazionale associata alla diffusione delle tecnologie di intelligenza artificiale. Secondo le stime riportate, nel 2022 il consumo energetico globale dei data center è stato pari a circa 460 terawattora (TWh).

Emergono informazioni importanti anche dagli studi condotti da Alex de Vries,

data scientist presso la banca centrale dei Paesi Bassi e dottore di ricerca alla Vrije Universiteit Amsterdam, che si occupa dell'analisi dei costi energetici associati alle tecnologie in fase di sviluppo. In uno dei lavori più recenti, egli ha stimato il potenziale impatto energetico di un'adozione su larga scala dell'intelligenza artificiale generativa, ipotizzando l'integrazione di sistemi come ChatGPT in ogni interrogazione effettuata tramite il motore di ricerca Google. Secondo tali stime, un simile scenario richiederebbe oltre 500.000 server Nvidia A100 HGX, corrispondenti a circa 4,1 milioni di unità di elaborazione grafica (GPU). Considerando una potenza assorbita di circa 6,5 kW per ciascun server, il consumo elettrico complessivo ammonterebbe a circa 80 GWh al giorno, ovvero 29,2 TWh su base annua (vedi Tabella 2.3). Tenendo in considerazione gli attuali limiti tecnologici dell'hardware e del software disponibili, de Vries ritiene poco realistico uno scenario di adozione così estesa dell'IA, principalmente a causa degli elevati costi economici e della limitata disponibilità di infrastrutture computazionali adeguate: fattori di natura tecnologica ed economica potrebbero costituire un freno significativo alla diffusione pervasiva dell'intelligenza artificiale e, di conseguenza, alla crescita del relativo consumo energetico. [7]

Parametro	**Valore**
Numero di server Nvidia A100 HGX	> 500.000
Numero totale di GPU	$\approx$ 4,1 milioni
Potenza assorbita per server	6,5 kW
Consumo elettrico totale giornaliero	$\approx$ 80 GWh/giorno
Consumo elettrico totale annuo	$\approx$ 29,2 TWh/anno

Tabella 2.3: Stima del consumo energetico di un'infrastruttura basata su server Nvidia A100 HGX

2.3 Dispendio energetico a confronto

2.3.1 Training ed inferenza

Un acceleratore di intelligenza artificiale è un dispositivo hardware specializzato progettato appositamente per incrementare l'efficienza computazionale e la velocità di esecuzione dei processi di addestramento (training) e di inferenza (inference) dei modelli di intelligenza artificiale, in particolare delle reti neurali profonde (deep learning). Tali dispositivi risultano ottimizzati per l'esecuzione parallela di operazioni matematiche caratterizzate da elevata intensità computazionale, quali, per esempio, le moltiplicazioni matriciali. Grazie a un'architettura dedicata e ad una maggiore larghezza di banda della memoria, gli acceleratori di AI sono in grado di offrire prestazioni superiori rispetto alle CPU convenzionali e, in alcuni casi, anche rispetto alle GPU.
Ogni nuova generazione di hardware implica un incremento delle prestazioni di elaborazione, tuttavia, tale miglioramento è generalmente accompagnato da un aumento del consumo energetico rispetto alla generazione precedente. Di conseguenza, all'aumentare dei volumi di utilizzo complessivi, si osserva una crescita significativa della domanda totale di energia.

In particolare, l'incremento del consumo energetico delle GPU, pari a circa il 75%, corrisponde ad un arco temporale relativamente breve (due anni) ossia il tempo di una singola generazione di sviluppo di nuovi modelli di GPU.
Un aspetto fondamentale per l'analisi dell'impatto energetico dei sistemi di intelligenza artificiale consiste nella distinzione tra la fase di addestramento di un modello e la sua successiva distribuzione e utilizzazione da parte degli utenti (inferenza).

- L'addestramento rappresenta il processo mediante il quale il modello apprende partendo da grandi quantità di dati e attraverso l'ottimizzazione iterativa di un elevato numero di parametri. Tale fase richiede l'esecuzione di un'enorme quantità di operazioni matematiche risultando quindi estremamente intensiva dal punto di vista computazionale ed energetico. Di conseguenza, il consumo di energia associato all'addestramento è significativamente superiore rispetto a quello delle attività ordinarie di un data center. Infatti, l'addestramento del modello GPT-3, che ha richiesto circa 1.287 MWh di energia elettrica, valore paragonabile al consumo annuo di circa 350–360 famiglie europee. Per quanto riguarda GPT-4 invece, le stime non ufficiali disponibili in letteratura e in analisi di settore variano da circa 1.700 MWh fino a oltre 7.000 MWh (vedi Figura 2.3), con proiezioni che suggeriscono valori ancora più elevati; i dati precisi, tuttavia, non sono di dominio pubblico.

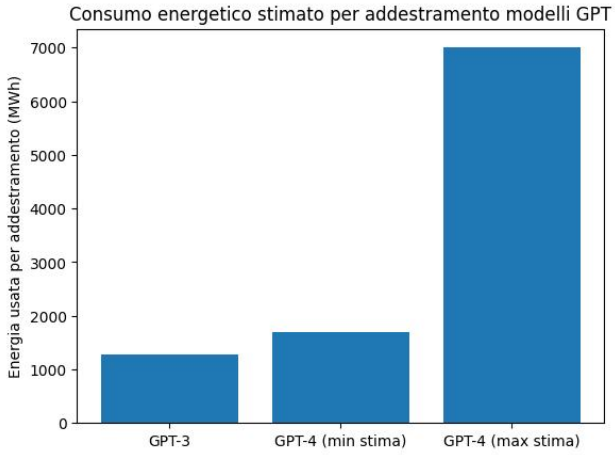

Figura 2.3: Stima del consumo energetico necessario per l'addestramento dei modelli GPT.

Al fine di contestualizzare tali ordini di grandezza, si consideri che lo streaming di un'ora di contenuti video su una piattaforma come Netflix richiede mediamente circa 0,8 kWh di energia. Per raggiungere un consumo energetico equivalente a quello stimato per l'addestramento di GPT-4, un singolo utente dovrebbe pertanto effettuare lo streaming ininterrottamente per un tempo dell'ordine di 2×10^6 ore.
I modelli linguistici di grandi dimensioni (Large Language Models, LLM) attualmente impiegati sono stati addestrati su insiemi di dati dell'ordine dei terabyte e presentano un numero di parametri pari o superiore a 175 miliardi.

Al termine della fase di addestramento, tali modelli vengono distribuiti in ambienti di produzione e avviano la fase di inferenza, durante la quale generano output in risposta a nuovi dati di ingresso.

- L'inferenza, nell'ambito dell'intelligenza artificiale, viene definita come il processo tramite il quale un modello utilizza le informazioni apprese durante l'addestramento per effettuare previsioni, prendere decisioni o risolvere specifici problemi. In un contesto applicativo, un modello di machine learning addestrato su un determinato dataset esegue l'inferenza per classificare nuove immagini, tradurre testi o stimare il valore futuro di una grandezza di interesse. Dunque, l'inferenza rappresenta l'applicazione operativa delle capacità del modello all'elaborazione di dati precedentemente non osservati.
  Nel caso di sistemi conversazionali quali ChatGPT, la fase di inferenza coincide con la generazione in tempo reale delle risposte alle interrogazioni degli utenti.
  Nell dibattito sulla sostenibilità ambientale dell'intelligenza artificiale, l'inferenza ha ricevuto un'attenzione relativamente minore rispetto alla fase di addestramento. Tuttavia, evidenze recenti suggeriscono che la fase di inferenza possa comunque contribuire in modo significativo ai costi complessivi, sia economici sia energetici, associati all'impiego di un modello di intelligenza artificiale. In altri termini, il processo attraverso il quale il modello applica le conoscenze acquisite è anche quello che può risultare maggiormente energivoro.

Secondo un'analisi condotta da SemiAnalysis, società indipendente di ricerca e analisi di settore, il supporto operativo di ChatGPT avrebbe richiesto l'impiego di 3.617 server NVIDIA HGX A100, per un totale di 28.936 GPU, corrispondenti a una domanda energetica stimata pari a circa 564 MWh al giorno. Tale valore risulta comparabile, e in alcuni casi superiore, ai 1.287 MWh stimati per la fase di addestramento di GPT-3. Ulteriori indicazioni in questa direzione provengono da Google, che ha riportato come circa il 60% del consumo energetico associato alle applicazioni di intelligenza artificiale nel periodo 2019–2021 sia attribuibile alla fase di inferenza. [13]

2.4 Deep Learning

Il Deep Learning, sottoinsieme del Machine Learning, si è affermato come una delle tecnologie più rilevanti e trasformative nel campo dell'intelligenza artificiale, ottenendo risultati importanti in molte applicazioni, dalla computer vision al natural language processing, fino ai sistemi di guida autonoma ed oltre. L'efficacia di tali modelli si fonda su una combinazione di principi teorici e sviluppi tecnologici che, interagendo tra loro, ne determinano l'elevata capacità espressiva e predittiva.
Un elemento centrale ed alla base del successo del Deep Learning è la capacità di apprendere rappresentazioni gerarchiche dei dati. Le reti neurali profonde, costituite da molteplici strati di elaborazione, sono in grado di estrarre caratteristiche a diversi livelli di astrazione. Negli strati iniziali vengono tipicamente individuate strutture semplici come trame nel caso delle immagini, mentre gli strati più profondi apprendono rappresentazioni sempre più complesse, fino a riconoscere oggetti, pattern semantici o relazioni di alto livello. Questo meccanismo multilivello consente ai

modelli di sviluppare una rappresentazione ricca e articolata dei dati, in modo concettualmente analogo ai processi cognitivi umani che integrano informazioni semplici in costrutti più complessi.
Un ulteriore fattore determinante per l'efficacia del Deep Learning è la disponibilità di grandi quantità di dati: dataset ampi e diversificati forniscono una base informativa più ricca per l'apprendimento delle strutture sottostanti ai fenomeni osservati, contribuendo così a migliorare la capacità di generalizzazione dei modelli. L'elevata dimensionalità dei dati e il numero di parametri coinvolti nelle reti profonde rendono infatti l'abbondanza di dati un requisito essenziale per un addestramento efficace.
Inoltre, la possibilità di addestrare modelli di Deep Learning su larga scala è stata fortemente agevolata dai progressi nell'hardware computazionale, in particolare dallo sviluppo di architetture dedicate come le GPU e le TPU. Queste unità di elaborazione offrono elevate capacità di calcolo parallelo, particolarmente adatte alle operazioni matriciali che caratterizzano l'addestramento delle reti neurali. La riduzione significativa dei tempi di addestramento ha quindi accelerato il ciclo di sviluppo e sperimentazione dei modelli ed ha anche reso praticabile l'esplorazione di architetture sempre più profonde e complesse, estendendo ulteriormente i limiti delle prestazioni raggiungibili dal Deep Learning. [14]

2.5 Verso un futuro sostenibile dell'intelligenza artificiale

La sostenibilità dell'intelligenza artificiale è una sfida centrale per il suo sviluppo futuro, in particolare alla luce del crescente impatto energetico e ambientale associato alle infrastrutture di calcolo. Una prima strategia per tentare di contenere tale impatto consiste nel favorire il riutilizzo di modelli generativi già esistenti, limitando così la creazione e l'addestramento di nuovi modelli. La fase di training, per esempio, richiede ingenti quantità di energia elettrica e di risorse computazionali: il riutilizzo e l'adattamento di modelli pre-addestrati (fine-tuning) consentirebbe dunque di ridurre il fabbisogno energetico complessivo.
Un ulteriore ambito di intervento riguarda l'adozione di metodi computazionali ad elevata efficienza energetica. Infatti, le architetture di calcolo come le CPU presentano consumi tipici dell'ordine di alcune decine di watt (circa 70 W), mentre le GPU ad alte prestazioni possono raggiungere potenze dell'ordine di alcune centinaia di watt (fino a circa 400 W). Al contrario, dispositivi a bassissimo consumo, quali microcontrollori o unità di elaborazione dedicate all'edge computing, operano con potenze dell'ordine delle centinaia di microwatt, risultando fino a mille volte più efficienti dal punto di vista energetico.
L'elaborazione locale dei dati (on-device o edge computing) consente una riduzione significativa del consumo energetico e della latenza evitando la trasmissione continua verso data center remoti e rappresentando una direzione promettente per lo sviluppo di applicazioni di IA sostenibili.
Un grande salto di qualità verso la sostenibilità dell'intelligenza artificiale può essere raggiunto solo attraverso una maggiore trasparenza nella misurazione e nella comunicazione dei consumi energetici: diventa essenziale quantificare, tracciare e rendere pubblici i dati relativi all'energia impiegata dai sistemi di IA e alle emissioni di CO_2

associate al loro utilizzo.
In risposta a questa esigenza, stanno emergendo nuovi indicatori nell'ambito dei criteri ESG (Environmental, Social and Governance), tra cui il Carbon Usage Effectiveness (CUE), che affianca il più tradizionale Power Usage Effectiveness (PUE) introducendo un parametro specificamente dedicato alla valutazione delle emissioni di anidride carbonica oltre che della potenza assorbita.
Inoltre, alcune aziende hanno iniziato a rendere disponibili informazioni dettagliate sul consumo energetico associato a singoli modelli di intelligenza artificiale o a unità funzionali standardizzate, quali il consumo per milione di inferenze eseguite. Si sta così facendo un passo significativo verso una maggiore responsabilizzazione del settore e una valutazione più consapevole dell'impatto ambientale delle tecnologie di IA.

È tuttavia importante sottolineare che l'intelligenza artificiale non costituisce esclusivamente una fonte di consumo energetico, ma può anche diventare uno strumento efficace per la riduzione dei consumi stessi. Nel settore energetico, ad esempio, sistemi basati sull'IA vengono impiegati per ottimizzare la gestione delle reti elettriche e migliorare la previsione della produzione da fonti rinnovabili, contribuendo ad una maggiore stabilità ed efficienza del sistema. In ambito industriale invece, tecniche di manutenzione predittiva supportate dall'intelligenza artificiale consentono di ridurre i tempi di inattività degli impianti e gli sprechi energetici associati, mentre negli edifici, algoritmi avanzati di controllo regolano in tempo reale sistemi di climatizzazione e di illuminazione, con riduzioni dei consumi energetici che possono anche raggiungere valori dell'ordine del 30%.
Più in generale, l'intelligenza artificiale può contribuire all'ottimizzazione della gestione delle risorse naturali, al miglioramento delle previsioni dei consumi energetici e al monitoraggio in tempo reale delle emissioni di CO_2. Tali benefici, tuttavia, possono concretizzarsi solo con l'adozione di tecnologie e architetture computazionali orientate alla riduzione dei consumi, ma anche all'integrazione di fonti energetiche rinnovabili all'interno delle infrastrutture dei data center e all'implementazione di sistemi di monitoraggio in grado di quantificare l'impatto ambientale di ciascun algoritmo. [13]

Capitolo 3

Scenari applicativi tecnico-scientifici

L'applicazione delle intelligenze artificiali trova particolare interesse nel campo della ricerca, soprattutto in ambito scientifico. Un riconoscimento importante è giunto nel 2024 con l'assegnazione di premi Nobel strettamente connessi allo sviluppo dell'IA. Il Premio Nobel per la Chimica 2024 è stato conferito a Demis Hassabis e John M. Jumper per lo sviluppo di un modello IA in grado di prevedere con elevata accuratezza la struttura tridimensionale delle proteine. Nel campo della Fisica invece, il Premio Nobel 2024 è stato assegnato a John J. Hopfield e Geoffrey E. Hinton per "Le scoperte e invenzioni fondamentali che consentono l'apprendimento automatico con reti neurali artificiali". In particolare, Hopfield introdusse nel 1982 il modello di rete neurale che porta il suo nome.
Oltre ai successi applicativi, un'ambizione dell'intelligenza artificiale risiede nella sua potenziale capacità di individuare nuove leggi fisiche o di suggerire possibili direzioni di ricerca. Grazie alla possibilità di analizzare grandi moli di dati e di identificare correlazioni non immediatamente evidenti, l'IA si configura come uno strumento promettente sia nell'analisi sperimentale sia nella simulazione computazionale di sistemi fisici complessi. [15]

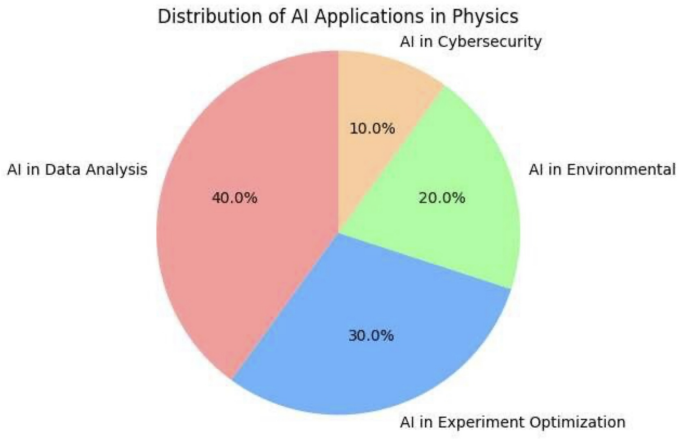

Figura 3.1: Distribuzione delle applicazioni AI nella fisica. [16]

Come mostrato in Figura 3.1, le principali applicazioni dell'intelligenza artificiale nella fisica possono essere suddivise in diverse aree: il 40% riguarda l'analisi dei dati sperimentali, particolarmente rilevante nella fisica delle alte energie; il 30% è legato all'ottimizzazione degli esperimenti, per migliorare la configurazione degli apparati

sperimentali e dei parametri di misura; il 20% delle applicazioni riguarda la fisica ambientale per il monitoraggio di sistemi ambientali complessi, come lo studio del cambiamento climatico; ed infine, circa il 10% delle applicazioni riguarda la sicurezza informatica per proteggere dati sperimentali sensibili ed infrastrutture di ricerca. [16]

3.1 Fisica ed Intelligenza artificiale

Nella ricerca in fisica, alcune delle principali sfide contemporanee riguardano proprio la gestione e l'interpretazione dei vasti volumi di dati prodotti dalle grandi collaborazioni scientifiche internazionali, la significativa riduzione dei tempi di analisi e, di conseguenza, i costi computazionali e operativi. In questo contesto, gli algoritmi di machine learning sono ormai strumenti preziosi ed essenziali per l'identificazione di segnali fisicamente rilevanti all'interno dei dati sperimentali, consentendo un'accelerazione significativa dei processi di analisi rispetto ai metodi tradizionali.
Il machine learning è ormai ampiamente utilizzato anche nella fisica delle particelle dove contribuisce allo studio e all'interpretazione dei processi fondamentali, ma si stanno aprendo prospettive interessanti per l'applicazione dell'IA generativa a problemi di natura più teorica. Tali tecniche di Machine Learning trovano applicazione, ad esempio, nel riconoscimento delle particelle prodotte in collisioni ad altissima energia, offrendo prestazioni superiori in termini di efficienza e accuratezza. Inoltre, l'apprendimento automatico permette di sviluppare modelli fisici sempre più complessi, rendendo le simulazioni numeriche più rapide e dettagliate.

Le basi teoriche dell'intelligenza artificiale affondano le proprie radici nella fisica statistica e nella teoria dei sistemi complessi. Sebbene i primi modelli teorizzati risalgano al secondo dopoguerra, è solo verso la fine del XX secolo che le applicazioni pratiche dell'IA diventano realmente fattibili, grazie ai progressi dell'informatica e alla crescente disponibilità di potenza di calcolo. Già a partire dagli anni Ottanta, i fisici hanno iniziato ad impiegare reti neurali e algoritmi di apprendimento automatico nell'analisi dei dati sperimentali. In fisica medica, ad esempio, l'IA trova applicazione in un ampio spettro di attività che spaziano dalla diagnostica per immagini ai gemelli digitali fino allo sviluppo di approcci di medicina di precisione. Tali innovazioni hanno svolto un ruolo rilevante anche nella determinazione degli elementi della matrice di Cabibbo–Kobayashi–Maskawa fondamentale per la comprensione della violazione di simmetria CP nei quark e nella scoperta del bosone di Higgs, che spiega l'origine della massa delle particelle elementari. [15]

3.1.1 IA nella ricerca al CERN di Ginevra

Alla fine degli anni Ottanta, presso il CERN di Ginevra e il Fermilab, iniziò a diffondersi per l'analisi dei dati e il riconoscimento delle tracce di particelle pesanti l'uso del machine learning. Viene discusso, per esempio, l'impiego di reti neurali per l'identificazione dei quark b e c nei dati prodotti dall'esperimento DELPHI all'acceleratore LEP.
Nel 2015, durante i preparativi per un nuovo ciclo di acquisizione dati dell'accelera-

tore LHC, un gruppo di fisici del CERN iniziò ad esplorare l'uso di tecniche di deep learning per l'analisi dei dati sperimentali. I primi studi si basarono su architetture come AlexNet per il riconoscimento delle particelle, applicate sia ai dati di LHC sia ad esperimenti sui neutrini. A partire dal 2017, gli esperimenti ATLAS e CMS hanno adottato modelli neurali più adatti alla gestione di strutture di dati complesse, quali le graph neural networks e, più recentemente, i transformer, la stessa famiglia di architetture utilizzata nello sviluppo di ChatGPT. Questi modelli sono in grado di cogliere relazioni altamente non lineari tra i segnali, migliorando di un ordine di grandezza la capacità di identificare eventi rari.
Tali progressi hanno favorito la nascita, all'interno degli esperimenti di LHC, di una comunità dedicata all'intelligenza artificiale impegnata nell'applicazione del machine learning a problemi precedentemente affrontati con metodi classici: simulazione dei rivelatori, selezione automatica degli eventi di interesse e ricostruzione delle particelle.
L'IA si sta affermando come standard operativo nella fisica delle alte energie: durante il Run 3, l'esperimento CMS ha compiuto introdotto reti neurali profonde nei sistemi di selezione in tempo reale dei dati (trigger). A partire dal 2024, CMS ha inoltre avviato la creazione di un archivio dedicato agli eventi anomali, con l'obiettivo di individuare segnali riconducibili a nuovi fenomeni fisici non ancora osservati. In questo scenario, l'intelligenza artificiale non si limita più a supportare l'analisi, ma diventa un elemento attivo della ricerca di frontiera. [15]

3.1.2 IA ed onde gravitazionali

La rivelazione delle onde gravitazionali costituisce uno dei campi più innovativi della fisica contemporanea. Dopo le prime scoperte, questi segnali sono diventati fondamentali per lo studio dell'universo in chiave astrofisica e cosmologica. Tuttavia, quelli registrati dagli interferometri LIGO, Virgo e KAGRA sono estremamente deboli e spesso immersi nel rumore strumentale. Dunque, la classificazione rapida e affidabile dei segnali è cruciale e poiché i dati sono rappresentabili come immagini tempo-frequenza, l'impiego di reti neurali convoluzionali si è dimostrato particolarmente efficace. [15]
Inoltre, nel settembre 2024 è stato proposto il Deep Loop Shaping, un nuovo metodo di intelligenza artificiale basato su machine learning e appositamente progettato per migliorare il controllo e la sensibilità dei rivelatori di onde gravitazionali di nuova generazione. Questo approccio risulta particolarmente importante nel contesto di interferometri laser estremamente sensibili, come Einstein Telescope, dove il controllo del rumore rappresenta un limite fondamentale alle prestazioni sperimentali.
Dal punto di vista metodologico, Deep Loop Shaping mostra come l'intelligenza artificiale possa essere impiegata non solo per l'analisi dei dati, ma anche per il controllo in tempo reale di sistemi fisici complessi, aspetto molto rilevante per la fisica sperimentale moderna, dove la crescente complessità degli apparati richiede strategie di controllo adattive.
Nei rivelatori di onde gravitazionali, il sistema di controllo ha il compito di mantenere gli specchi dell'interferometro in condizioni di stabilità estrema, sopprimendo vibrazioni ambientali e perturbazioni meccaniche. Tuttavia, un controllo troppo aggressivo può introdurre rumore di controllo, degradando la sensibilità del

rivelatore in specifiche bande di frequenza rappresentando così uno dei principali ostacoli al miglioramento della sensibilità degli interferometri futuri.
Deep Loop Shaping affronta questo problema superando i metodi di controllo lineare tradizionali. Il metodo infatti utilizza algoritmi di machine learning addestrati su simulazioni realistiche del rivelatore e del suo ambiente, riproducendo iterativamente il problema di controllo. In questo modo, l'algoritmo apprende una strategia di controllo ottimale capace di minimizzare il rumore complessivo del sistema, mantenendo comunque la stabilità richiesta per le misurazioni interferometriche. [17]
Applicazioni a interferometri esistenti come LIGO e Virgo indicano che tale tecnica potrebbe consentire l'osservazione di centinaia di eventi di onde gravitazionali aggiuntivi ogni anno con un livello di precisione superiore. I risultati dei test sperimentali indicano una riduzione del rumore ed un miglioramento del controllo del sistema compresi tra un fattore 30 e 100 rispetto alla strumentazione attuale, miglioramento che si traduce in un aumento della sensibilità e quindi in un incremento del volume di universo osservabile. [15]

In prospettiva, l'integrazione di tecniche di machine learning nei sistemi di controllo sarà un elemento chiave per sfruttare appieno il potenziale scientifico dei rivelatori di onde gravitazionali di prossima generazione, aprendo la strada a misure ancora più precise.

3.1.3 Machine Learning in astrofisica

Negli ultimi dieci anni, il machine learning (ML) si è affermato come uno strumento estremamente efficace in numerose applicazioni astronomiche. Tali tecniche, si sono dimostrate vantaggiose in particolare nell'affrontare problemi complessi per i quali i metodi analitici tradizionali risultano spesso poco efficienti o computazionalmente onerosi. Tra le applicazioni più rilevanti si trovano la stima dei redshift fotometrici, la determinazione dei parametri strutturali delle galassie a partire dai profili di luminosità, la classificazione di stelle, galassie e quasar sia su immagini sia su cataloghi osservativi, nonché l'identificazione e la modellizzazione di eventi di lente gravitazionale forte. Ulteriori ambiti di successo includono la ricostruzione della distribuzione di materia oscura (dark matter) nelle galassie e lo studio delle relazioni tra le proprietà delle galassie e quelle degli aloni di materia oscura che le ospitano.
Più recentemente, sono stati proposti approcci basati su algoritmi di ML addestrati su simulazioni cosmologiche per stimare il contenuto di materia oscura delle galassie. Le simulazioni cosmologiche, fondate su principi fisici ben definiti e su modelli realistici dei processi di feedback, producono popolazioni galattiche che riproducono in maniera realistica le distribuzioni osservate e le principali relazioni di scala. Per questo motivo rappresentano un ambiente ideale per l'addestramento di modelli di machine learning finalizzati all'inferenza delle proprietà della materia oscura.
Tra le metodologie più utilizzate vi sono le random forests (RF), un approccio di ensemble learning basato su alberi decisionali (Classification and Regression Trees). La predizione finale di una RF si ottiene mediando le predizioni di M alberi con la seguente formula, che rappresenta una schematizzazione semplificata del meccanismo

di aggregazione tipico dei metodi ensemble:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \tag{3.1}$$

dove $T_m(x)$ rappresenta la predizione del m-esimo albero decisionale sull'input x, e M è il numero totale di alberi nell'ensemble. Questa tecnica è utilizzata per ottenere stime di proprietà fisiche come la massa ed il raggio di mezza massa degli aloni, riducendo il rischio di overfitting grazie alla media delle predizioni dei singoli modelli.
Per l'addestramento dei modelli, invece, i dataset possono essere suddivisi in campioni globali o specializzati, a seconda del tipo di galassia (ellittiche, nane, tardive), con l'80% dei dati utilizzato per il training e il 20% per il test. Nel caso di applicazioni a dati osservativi, l'intero dataset simulato viene impiegato per l'addestramento.
L'efficacia dei modelli viene valutata mediante metriche standard che permettono di quantificare rigorosamente l'accuratezza e la precisione dei modelli. Dunque, il machine learning su dati simulati costituisce uno strumento potente per inferire proprietà della materia oscura partendo da osservabili galattici e fornendo informazioni complementari a quelle ottenibili tramite metodi analitici o pure osservazioni dirette.[18]

Nel grafico a barre in Figura 3.2, è possibile vedere come l'astrofisica presenti il valore più elevato nell'impatto dovuto all'IA, pari a 85, poiché si necessita dell'elaborazione grandi quantità di dati.
Inoltre, vengono mostrati anche gli impatti relativi ad altri settori della fisica. In particolare, nella fisica delle particelle il livello di impatto è pari a 75 poiché l'IA viene utilizzata per l'analisi dei dati provenienti dagli acceleratori di particelle e dagli esperimenti ad alte energie. Nella meccanica quantistica l'impatto è pari a 65, per l'ottimizzazione e la risoluzione di esperimenti quantistici. Infine, nelle scienze del clima, l'impatto corrisponde a 55: l'AI contrisce alla previsione dei modelli meteorologici e alla valutazione dei consumi energetici.

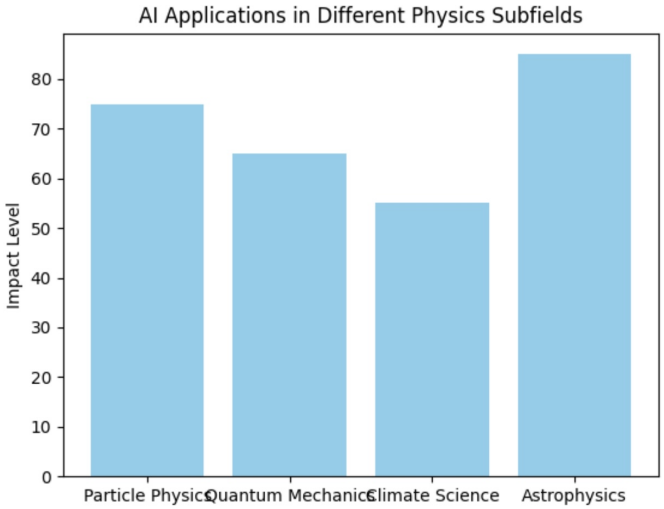

Figura 3.2: Applicazioni AI nei diversi campi di ricerca della fisica. [16]

3.2 Accuratezza ed efficienza dei dati

L'accuratezza dei dati rappresenta una dimensione centrale della qualità dell'informazione, insieme a completezza, coerenza, tempestività, unicità e validità. Dal punto di vista operativo, dati accurati costituiscono un prerequisito essenziale affinché i risultati di analisi siano attendibili ed affidabili per l'utilizzo in diversi ambiti applicativi. Tra questi si trovanp il miglioramento dell'efficienza operativa, il rispetto dei requisiti normativi, la qualità degli output dei sistemi di intelligenza artificiale ed una maggiore soddisfazione degli utenti.
Quando i dati sono affidabili e coerenti, i processi decisionali risultano maggiormente allineati agli indicatori chiave di prestazione (Key Performance Indicators, KPI), aumentando quindi l'efficacia complessiva delle strategie adottate. Al contrario, la presenza di dati errati o imprecisi compromette l'affidabilità delle decisioni e può generare effetti negativi a cascata sulle attività operative. Un ulteriore aspetto critico riguarda la conformità normativa: dati incompleti o inaccurati espongono le organizzazioni al rischio di violazioni delle normative e degli standard di settore.
L'accuratezza dei dati è strettamente legata con il concetto di integrità dei dati, anche se i due concetti non coincidono. Quest'ultima si riferisce alla capacità di preservare accuratezza, completezza e coerenza delle informazioni lungo l'intero ciclo di vita del dato, anche in presenza di trasferimenti tra sistemi differenti. Infatti, l'accuratezza rappresenta un elemento chiave dell'integrità, poiché garantisce che ciascun dato descriva correttamente il fenomeno reale a cui si riferisce.

La crescente complessità computazionale dei modelli di intelligenza artificiale introduce sfide rilevanti in termini di efficienza, costi operativi e impatto ambientale. L'ottimizzazione dell'efficienza risulta quindi cruciale non solo per il controllo dei costi, ma anche per la riduzione dell'impronta carbonica dell'IA, considerando che i data center sono responsabili di circa l'1% del consumo globale di energia elettrica.
L'efficienza di un modello di intelligenza artificiale può essere definita come la capacità di utilizzare in modo ottimale le risorse computazionali mantenendo un livello di accuratezza adeguato. Tale concetto va oltre la sola velocità di esecuzione ed include aspetti quali l'utilizzo della memoria, il consumo energetico e la compatibilità con specifiche architetture hardware.
Durante la fase di addestramento, l'efficienza riguarda principalmente la gestione delle risorse di calcolo e dei dati; nella fase di inferenza, invece, sono centrali la latenza e la rapidità di risposta. L'uso efficiente della memoria e dell'energia incide in modo significativo sia sui costi operativi sia sulla sostenibilità ambientale dei sistemi. [19]
Le prestazioni di efficienza tendenzialmente vengono valutate attraverso metriche come il throughput, la latenza, il grado di utilizzo di CPU e GPU e l'accuratezza del modello. Ulteriori fattori determinanti sono la scalabilità, l'efficienza temporale e le ottimizzazioni specifiche per l'hardware di destinazione. In contesto applicativo, l'ottimizzazione efficace richiede un bilanciamento attento tra il risparmio computazionale e le eventuali perdite di accuratezza che devono essere valutate in funzione dello scenario di utilizzo. Questo equilibrio è formalizzato nel concetto di efficient accuracy trade-off, che mira a individuare il punto ottimale in cui un modello garantisce buone prestazioni predittive senza risultare però eccessivamente oneroso dal

punto di vista computazionale.
Nelle applicazioni pratiche, in particolare nei sistemi in tempo reale, tale compromesso diventa ancora più critico: gli sviluppatori devono progettare modelli accurati, ma soprattutto sufficientemente leggeri da poter essere eseguiti su dispositivi con risorse limitate. Vengono dunque adottate tecniche come il model pruning, la quantizzazione e l'impiego di architetture leggere, che consentono di ridurre le dimensioni e la complessità dei modelli preservandone la capacità predittiva.
Un esempio significativo di questo approccio è rappresentato da YOLOv8, che dimostra come sia possibile bilanciare efficienza e accuratezza. Grazie ad una progettazione modulare e a strategie di addestramento ottimizzate, questo modello evidenzia che prestazioni elevate nel riconoscimento di oggetti non richiedono necessariamente di sacrificare velocità di esecuzione. [20]

3.3 Impronta energetica del cloud scientifico

In questi ultimi anni, nel settore del High-Performance Computing (HPC), è emerso l'impiego di acceleratori applicativi, in particolare delle GPU, per incrementare le prestazioni di codici caratterizzati da un'elevata intensità computazionale. Sono stati realizzati numerosi cluster HPC di grandi dimensioni dotati di GPU, come ad esempio il cluster Lincoln presso il National Center for Supercomputing Applications (NCSA). Oltre all'aumento delle prestazioni, è cresciuto di conseguenza anche l'interesse verso l'analisi del consumo energetico di tali infrastrutture, con l'obiettivo di comprendere ed ottimizzare il rapporto tra prestazioni e potenza assorbita.
Alcuni studi hanno affrontato il problema dell'utilizzo di strumenti necessari per la misurazione e l'analisi della potenza all'interno di cluster HPC potenziati da GPU, con lo scopo di fornire un profilo dettagliato del consumo energetico delle applicazioni eseguite. La disponibilità di queste informazioni è fondamentale per lo sviluppo di applicazioni energy-aware, per la progettazione di sistemi orientati al risparmio energetico e per l'implementazione di strategie di gestione delle risorse capaci di ottimizzare simultaneamente prestazioni computazionali e consumo di potenza.
Un ulteriore passo in questa direzione è rappresentato dalla proposta di modelli statistici per stimare il consumo energetico delle GPU. In particolare, sono stati sviluppati modelli di regressione basati sulla misura della potenza assorbita durante l'esecuzione di diversi benchmark e sull'identificazione dei blocchi funzionali della GPU coinvolti nei vari carichi di lavoro. Partendo da tali osservazioni, il modello statistico consente di predire il consumo energetico di una GPU target durante l'esecuzione di una specifica applicazione, offrendo uno strumento utile per la pianificazione e l'ottimizzazione dei carichi computazionali.
Invece, altri studi hanno proposto architetture server in grado di integrare il monitoraggio energetico in tempo reale dei singoli componenti del sistema, tra cui CPU, GPU, scheda madre e memoria. Queste architetture sono state validate sperimentalmente permettendo di acquisire dati specifici sul comportamento energetico del sistema. I risultati mostrano che l'utilizzo delle GPU risulta energeticamente più efficiente solo quando il guadagno prestazionale ottenuto supera una determinata soglia, evidenziando come l'accelerazione hardware non garantisca automaticamente

una riduzione del consumo energetico complessivo.
Inoltre, è importante citare anche lo sviluppo di framework hardware-software completi per l'analisi del consumo energetico delle applicazioni parallele su sistemi multi-core. Un esempio significativo è rappresentato dal framework PowerPack, che combina un'infrastruttura hardware basata su sonde per la misurazione della potenza assorbita dai singoli componenti (CPU, memoria, disco, ecc.) con un ambiente software in grado di raccogliere e correlare i dati energetici durante l'esecuzione delle applicazioni. Questo approccio consente un'analisi approfondita del consumo energetico fornendo strumenti essenziali per la progettazione di applicazioni HPC più efficienti e sostenibili.

Nei sistemi di High-Performance Computing accelerati da GPU, il monitoraggio accurato del consumo energetico richiede soluzioni hardware che risultino al contempo affidabili, poco invasive ed economicamente sostenibili. Un primo approccio efficace è rappresentato dall'impiego di dispositivi di monitoraggio della potenza a basso costo ed integrati con moduli di comunicazione wireless. Questi sistemi consentono la misura di tensione e corrente e la trasmissione periodica dei dati ad un nodo di raccolta, permettendo di distinguere il consumo energetico del nodo host da quello dell'unità di accelerazione GPU.
Tuttavia, soluzioni di questo tipo presentano alcune limitazioni, in particolare per quanto riguarda la frequenza di campionamento, la scalabilità e il supporto alle tensioni tipiche degli ambienti HPC. Per superare tali criticità, è possibile progettare unità di monitoraggio dedicate, realizzate sotto forma di Power Distribution Unit (PDU) intelligenti, che siano in grado di monitorare simultaneamente più linee di alimentazione. La corrente su ciascun canale viene misurata mediante trasformatori di corrente in corrente alternata, mentre un microcontrollore acquisisce i segnali analogici e calcola valori efficaci (RMS) accurati su finestre temporali caratterizzate da un'elevata densità di campionamento.
La possibilità di configurare i resistori di carico permette di adattare la sensibilità del sistema ai diversi profili di potenza, massimizzando la precisione delle misure senza però compromettere la sicurezza dell'hardware. Anche in caso di superamento del range di misura, il sistema continua ad operare segnalando condizioni di saturazione ed evitando quindi danni ai componenti. L'integrazione di protezioni elettriche standard garantisce un livello di affidabilità comparabile a quello delle PDU commerciali, mentre l'esportazione dei dati attraverso interfacce comuni come USB facilita l'integrazione con infrastrutture software già esistenti.
Dunque, un'architettura di monitoraggio energetico progettata appropriatamente consente di ottenere misure con una buona risoluzione temporale e con un errore contenuto, mantenendo anche costi ridotti. Questo approccio rende il monitoraggio energetico su larga scala una soluzione praticabile nei cluster HPC e costituisce la base per l'analisi dell'efficienza energetica, il confronto tra differenti configurazioni hardware e l'ottimizzazione delle applicazioni accelerate da GPU.

L'efficienza energetica delle applicazioni HPC accelerate da GPU viene analizzata grazie a dati di potenza raccolti ed articolati su tre elementi principali:

- Sensori: ogni sensore è identificato in modo univoco e associato al dispositivo monitorato e al proprio intervallo di acquisizione.

- Dati di potenza: comprendono i valori di tensione e corrente misurati, collegati al sensore e al relativo istante di campionamento.
- Dati dei job: contengono informazioni sui job che impiegano le risorse monitorate, inclusi orari di inizio e fine, utente e nome del job utili per generare grafici temporali del consumo energetico.

La misura fondamentale utilizzata per confrontare l'efficienza energetica tra versioni CPU-only e GPU-accelerate è il miglioramento delle prestazioni per watt definito come:

$$e = \frac{p_{\text{CPU}}}{p_{\text{GPU}}} \cdot s \tag{3.2}$$

dove

$$p_{\text{CPU}} : \text{ potenza media consumata durante l'esecuzione su CPU,} \tag{3.3}$$

$$p_{\text{GPU}} : \text{ potenza media consumata durante l'esecuzione accelerata su GPU,} \tag{3.4}$$

$$s = \frac{t_{\text{CPU}}}{t_{\text{GPU}}} \tag{3.5}$$

Il parametro s rappresenta il fattore di accelerazione (speedup), ossia il rapporto tra il tempo di esecuzione su CPU (t_{CPU}) ed il tempo di esecuzione su GPU (t_{GPU}).

La grandezza e rappresenta un indice relativo di efficienza energetica, che mette in relazione il miglioramento prestazionale con la potenza media assorbita dai due sistemi. Non misura direttamente l'energia consumata, ma confronta il rapporto tra prestazioni e potenza (performance per watt) nelle configurazioni CPU e GPU.
Questa formula consente di quantificare quante volte l'implementazione GPU risulti più efficiente dal punto di vista energetico rispetto alla versione CPU-only. È quindi necessario distinguere tra potenza media, ossia il tasso istantaneo di consumo energetico durante l'esecuzione; tempo di esecuzione, che determina la durata del calcolo; ed energia totale, ottenuta come prodotto tra potenza e tempo.
Inoltre, è importante che la raccolta dei dati escluda le misure durante le fasi di avvio e terminazione dell'applicazione, così da ottenere valori medi rappresentativi di una simulazione a regime. I dati di potenza vengono mediati su più campioni e filtrati per garantire una valutazione accurata e per escludere le fasi iniziali e finali delle simulazioni in cui CPU e GPU non sono ancora a regime. [21]

3.3.1 Esempi di applicazioni scientifiche

- **Simulazioni di dinamica molecolare**

  In un sistema virale di circa un milione di atomi, l'accelerazione GPU riduce il tempo di un singolo timestep da 6.6 s a 1.1 s, con fattore di accelerazione

  $$s = \frac{6.6}{1.1} = 6.$$

  La potenza media della versione CPU-only è $p_{\text{CPU}} = 316\,\text{W}$, mentre la configurazione CPU+GPU consuma $p_{\text{GPU}} = 681\,\text{W}$.

  Il miglioramento delle prestazioni per watt risulta quindi:

  $$e = \frac{316}{681} \cdot 6 \approx 2.78.$$

La versione GPU risulta quasi tre volte più efficiente dal punto di vista energetico rispetto alla versione solo CPU.

- **Visualizzazione e analisi molecolare**

  Per il calcolo del potenziale elettrostatico su una traiettoria di circa 685 000 atomi, il tempo di esecuzione passa da 1465 s a 57.5 s:

  $$s = \frac{1465}{57.5} \approx 25.5.$$

  La potenza media è $p_{\text{CPU}} = 299\,\text{W}$ e $p_{\text{GPU}} = 742\,\text{W}$.

  Il miglioramento per watt è:

  $$e = \frac{299}{742} \cdot 25.5 \approx 10.48.$$

  In questo caso, l'accelerazione GPU porta ad un incremento di efficienza energetica superiore a dieci volte.

- **Simulazioni quantistiche Monte Carlo**

  In simulazioni di un cristallo di diamante con 512 elettroni, la versione GPU raggiunge un fattore di accelerazione

  $$s = 61.5.$$

  Il consumo medio è $p_{\text{CPU}} = 314\,\text{W}$ e $p_{\text{GPU}} = 853\,\text{W}$.

  L'efficienza energetica risultante è:

  $$e = \frac{314}{853} \cdot 61.5 \approx 22.6.$$

  La GPU rende la simulazione oltre venti volte più efficiente in termini di prestazioni per watt.

- **Calcoli di cromodinamica quantistica**

  Per una lattice QCD 4D, il tempo di esecuzione si riduce da 77324 s a 3881 s:

  $$s = \frac{77324}{3881} \approx 19.9.$$

  La potenza media passa da $p_{\text{CPU}} = 225\,\text{W}$ a $p_{\text{GPU}} = 554\,\text{W}$.

  Il miglioramento per watt è:

  $$e = \frac{225}{554} \cdot 19.9 \approx 8.1.$$

  Anche in questo caso, l'accelerazione GPU determina un notevole incremento dell'efficienza energetica.

Questa analisi mostra come l'efficienza energetica dipende dall'aggiunta di GPU ma soprattuttto dal fattore di accelerazione. Le applicazioni con speedup elevato ottengono risparmi energetici importanti al contrario delle accelerazioni minori che invece non garantiscono miglioramenti per watt.
Il confronto tra CPU-only e CPU+GPU necessita di un monitoraggio dettagliato del consumo separato di host e GPU, sottolineando così l'importaza di misure affidabili e precise.

3.3.2 Analisi dell'efficienza energetica

L'efficienza energetica di un'applicazione HPC accelerata da GPU può essere modellata in funzione del fattore di accelerazione s e del consumo di potenza del sistema. Indicando con P_{host} la potenza media della configurazione CPU-only e con $P_{\text{host+GPU}}$ la potenza media della configurazione accelerata CPU+GPU, il miglioramento delle prestazioni per watt può essere espresso come

$$e = \frac{P_{\text{host}}}{P_{\text{host+GPU}}} \cdot s. \tag{3.6}$$

Questa formula consente di valutare quanto l'impiego di GPU migliori complessivamente l'efficienza energetica tenendo anche in considerazione il consumo energetico dell'hardware dell'acceleratore. Applicazioni scientifiche GPU-accelerate possono risultare fino a un ordine di grandezza più efficienti dal punto di vista energetico. Dunque, è presente una forte dipendenza dal fattore di accelerazione per quanto riguarda il beneficio energetico: valori elevati di s producono risparmi energetici significativi, mentre accelerazioni modeste possono non compensare l'aumento di potenza assorbita dal sistema accelerato.

In Figura 3.3, è possibile osservare come un consumo totale maggiore implichi un miglioramento per watt minore, ossia all'aumentare del consumo di potenza di CPU+GPU, il beneficio in termini di prestazioni per watt diminuisce.
Inoltre, per riuscire ad ottenere un significativo miglioramento energetico, il fattore di accelerazione deve superare un valore soglia: accelerazioni minime non garantiscono risparmio energetico, mentre accelerazioni elevano comportano guadagni significativi.
È doveroso specificare però che monitorare singolarmente le GPU in sistemi multi-GPU può essere complesso: il consumo delle GPU inattive e dei core CPU non utilizzati viene incluso nelle misure complessive. Tuttavia, queste misure rappresentano il consumo reale del sistema durante l'esecuzione dell'applicazione, indipendentemente dall'efficienza di utilizzo delle risorse. Una configurazione più accurata e flessibile consente di ottenere valori più precisi di efficienza energetica, migliorando l'analisi comparativa tra diverse applicazioni o configurazioni hardware. [21]

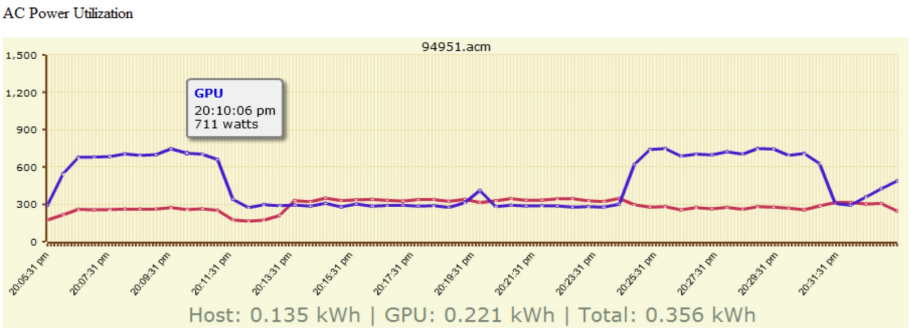

Figura 3.3: Andamento del consumo energetico per ogni dispositivo nel tempo. [21]

Capitolo 4

Intelligenza artificiale nella vita quotidiana

Le tecnologie ed i sistemi algoritmici che utilizzano l'intelligenza artificiale sono oggi profondamente integrati in numerosi ambiti della vita privata e pubblica, incluse le relazioni sociali, le transazioni economiche, la ricerca scientifica e la gestione delle produzioni. In particolare, gli algoritmi di apprendimento automatico hanno un ruolo rilevante nella mediazione delle interazioni tra utenti e piattaforme digitali, influenzando anche in modo significativo l'accesso alle informazioni, le scelte individuali ed i comportamenti collettivi.
Le piattaforme digitali operano attraverso sistemi di raccolta, aggregazione ed analisi di grandi volumi di dati eterogenei (big data), provenienti sia da interazioni esplicite degli utenti sia da banali segnali comportamentali impliciti. Tali dati vengono poi utilizzati per costruire modelli predittivi dei profili individuali, personalizzando i contenuti mostrati nei feed informativi. Questo processo è alla base dei sistemi di raccomandazione impiegati, ad esempio, nei social network, nei servizi di streaming e nelle piattaforme pubblicitarie.
L'aumento della quantità dei dati disponibili migliora le prestazioni predittive incrementando l'efficacia dei contenuti raccomandati in termini di engagement e permanenza sulla piattaforma.
Tuttavia, questa ottimizzazione introduce anche alcuni effetti collaterali rilevanti dal punto di vista socio-tecnico, ad esempio, i sistemi algoritmici possono amplificare dinamiche di polarizzazione del dibattito pubblico influenzando i processi di formazione dell'opinione individuale e collettiva.
Dunque, l'azione degli algoritmi non si limita ad una funzione di supporto decisionale, ma può configurarsi come un vero e proprio meccanismo di orientamento dell'informazione e, potenzialmente, di indirizzamento di massa delle preferenze e delle opinioni.
Questi aspetti evidenziano la necessità di un'analisi critica dell'impatto degli algoritmi sul piano etico, sociale e democratico, nonché dell'adozione di principi di trasparenza, responsabilità e governance nei sistemi di intelligenza artificiale impiegati su larga scala.

4.1 Algoritmi e servizi digitali quotidiani

4.1.1 Sistemi di raccomandazione

Un sistema di raccomandazione è un algoritmo di filtraggio dell'informazione progettato per stimare le preferenze di un utente con l'obiettivo di suggerire contenuti rilevanti in modo personalizzato. Tali sistemi rappresentano una componente fondamentale delle moderne applicazioni di intelligenza artificiale e sono ampiamente utilizzati per gestire l'elevata quantità di informazioni disponibili online, riducendo il sovraccarico informativo e supportando il processo decisionale dell'utente.
I sistemi di raccomandazione operano analizzando Big Data come cronologia, pattern di navigazione, interazioni implicite (click, tempo di visualizzazione) ed esplicite (valutazioni, recensioni) e, attraverso l'individuazione di correlazioni e regolarità statistiche nei dati, consentono di aumentare il coinvolgimento degli vari utenti.

Dal punto di vista implementativo ed algoritmico, i motori di raccomandazione si basano prevalentemente su tecniche di Machine Learning, che possono essere ricondotte a tre principali categorie (vedi Figura 4.1):

- Filtraggio collaborativo (Collaborative Filtering)

  È un approccio fondato sull'ipotesi che utenti con comportamenti simili nel passato tenderanno ad avere preferenze simili anche in futuro. Le raccomandazioni vengono generate identificando analogie tra utenti (user-based) o tra elementi (item-based) sulla base delle interazioni osservate. Queste interazioni vengono poi formalizzate attraverso una matrice di interazione utente–articolo, i cui elementi rappresentano feedback espliciti (valutazioni, like) o impliciti. Ad esempio, se due utenti hanno valutato positivamente uno stesso contenuto, il sistema può inferire una similarità di preferenze e suggerire elementi apprezzati da uno all'altro.

  I metodi di collaborative filtering si distinguono ulteriormente in due sottocategorie a seconda della strategia utilizzata per stimare tali similitudini: metodi memory-based e metodi model-based.

  - Metodi memory-based: operano direttamente sulla matrice di interazione utente–articolo, senza costruire un modello esplicito. Si distinguono in approcci user-based dove ogni utente è rappresentato come un vettore delle proprie interazioni con gli articoli e la raccomandazione per un utente target viene generata selezionando gli articoli più rilevanti tra quelli apprezzati dal suo vicinato, ovvero l'insieme degli utenti più simili.
    Oppure in approcci item-based dove invece sono gli articoli ad essere rappresentati in base alle interazioni ricevute dagli utenti, le raccomandazioni vengono prodotte suggerendo articoli simili a quelli con cui l'utente ha interagito positivamente in passato.
  - Metodi model-based: assumono che le interazioni osservate tra utenti e articoli siano generate da un modello latente sottostante imparato dai

dati. Invece di lavorare direttamente sulla matrice di interazione, questi approcci apprendono una rappresentazione compatta e astratta delle preferenze consentendo di migliorare la capacità di generalizzazione del sistema.

- Filtraggio basato sui contenuti (Content-Based Filtering)

  In questo caso, le raccomandazioni vengono prodotte analizzando le caratteristiche intrinseche degli elementi e confrontandole con il profilo dell'utente sfruttando informazioni descrittive aggiuntive (su utente e/o articoli). Il sistema suggerisce contenuti simili a quelli precedentemente apprezzati con l'obiettivo di costruire un modello che apprenda la relazione tra le feature degli utenti e quelle degli articoli, spiegando così la matrice di interazione osservata. Questo rende l'approccio content-based perfetto per contesti in cui le interazioni storiche sono limitate, sebbene possa avere una minore capacità di esplorazione rispetto ai metodi collaborativi.

- Modelli ibridi (Hybrid Recommender Systems)

  I sistemi ibridi combinano filtraggio collaborativo e basato sui contenuti al fine di mitigare i limiti dei singoli approcci, come il problema del cold start (assenza di dati storici per nuovi utenti o nuovi elementi) o la scarsa capacità di generalizzazione in contesti poco densi di interazioni.
  Integrando informazioni comportamentali e descrittive, tali sistemi migliorano l'accuratezza delle raccomandazioni, la scarsità di dati e la capacità di adattamento a nuovi utenti o contenuti.
  [22]

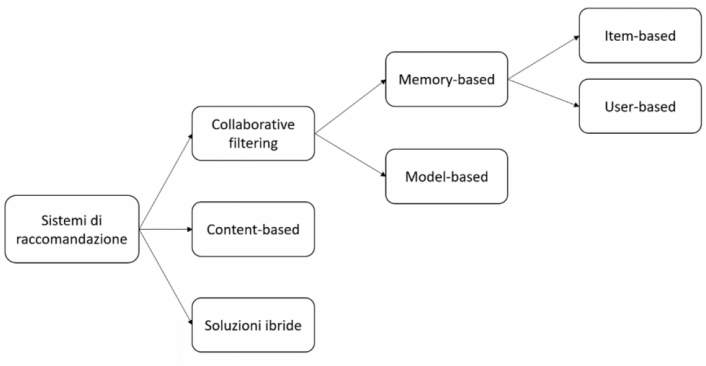

Figura 4.1: Schematizzazione dei diversi sistemi di raccomandazione. [22]

È importante distinguere i sistemi di raccomandazione dalle tecnologie abilitanti che ne costituiscono invece solo una parte:

- Ricerca vettoriale (Vector Search)

  È una tecnica di recupero dell'informazione che consente di individuare gli elementi più simili a una query all'interno di uno spazio vettoriale, sulla base di misure di distanza o somiglianza. Nonostante la ricerca vettoriale sia frequentemente utilizzata nei sistemi di raccomandazione, questi ultimi includono

una logica più ampia che comprende la profilazione degli utenti e l'adattamento dinamico delle raccomandazioni.

- Ricerca semantica (Semantic Search)

  A differenza dei metodi basati su semplici corrispondenze comportamentali, la ricerca semantica mira a catturare il significato e l'intento sottostante a una query. Nei sistemi di raccomandazione può essere impiegata per interpretare il contesto delle interazioni dell'utente e migliorare la pertinenza dei contenuti suggeriti, soprattutto in ambienti informativi più complessi. [23]

Vengono applicati i sistemi di raccomandazione anche in molti altri contesti tra cui siti di informazione e portali di notizie, videogiochi e piattaforme di intrattenimento interattivo, sistemi di supporto alla ricerca, social media e piattaforme di condivisione di contenuti, sistemi di supporto alle decisioni nel trading finanziario.
L'integrazione di un sistema di raccomandazione all'interno di un sito web o di un'applicazione software offre numerosi benefici come, ad esempio, l'incremento delle vendite, grazie alla proposta di offerte personalizzate, un miglioramento dell'esperienza utente rendendo la fruizione dei contenuti più pertinente e mirata o l'aumento del tempo di permanenza sulla piattaforma favorito dalla rilevanza delle raccomandazioni.
Uno studio recente condotto da Epsilon ([24]) evidenzia che il 90% dei consumatori considera la personalizzazione un aspetto positivo e circa l'80% dichiara di essere più propenso a interagire con un'azienda che offre esperienze personalizzate. Inoltre, lo stesso studio mostra che tali utenti hanno una probabilità dieci volte maggiore di diventare VIP customers effettuando più di quindici acquisti all'anno.
Dunque, i sistemi di raccomandazione, erogando contenuti personalizzati e adattandosi alle preferenze individuali degli utenti, contribuiscono in modo significativo sia al successo economico delle aziende sia alla qualità dell'esperienza offerta agli utenti. [25]

4.1.2 Assistenti vocali

Un assistente di intelligenza artificiale è un'applicazione software che utilizza tecniche di IA per comprendere il linguaggio naturale, interpretare comandi dell'utente ed eseguire azioni. Questi sistemi sono progettati per interagire in modo naturale con l'essere umano riducendo la complessità dell'accesso ai servizi digitali.
Gli assistenti AI si basano su tecnologie avanzate quali l'elaborazione del linguaggio naturale (Natural Language Processing), l'apprendimento automatico (Machine Learning) e l'analisi dei dati grazie alle quali sono in grado di apprendere progressivamente dalle interazioni con gli utenti, migliorando la qualità delle risposte, la capacità di contestualizzazione e l'anticipazione delle esigenze, fino a fornire suggerimenti e servizi personalizzati.
È opportuno distinguere tra assistenti vocali e smart speaker:

- l'assistente vocale è il componente software responsabile del riconoscimento vocale, dell'interpretazione dei comandi e della generazione delle risposte;

- lo smart speaker è il dispositivo hardware che integra un assistente vocale e consente l'interazione tramite comandi vocali.

Gli assistenti vocali hanno trasformato il modo in cui gli utenti interagiscono con la tecnologia nel quotidiano. Le principali funzionalità comprendono il controllo di dispositivi intelligenti, come sistemi di illuminazione, climatizzazione e sicurezza domestica, la riproduzione multimediale con avvio e gestione di contenuti audio e video su diversi dispositivi, e il supporto alla produttività personale attraverso promemoria, sveglie, calendari e appuntamenti. Inoltre, offrono accesso rapido alle informazioni permettendo interrogazioni su meteo, traffico, notizie e conoscenze generali, e facilitano il commercio elettronico con l'utilizzo di servizi digitali che consentono l'acquisto di prodotti e l'accesso a servizi tramite interazione vocale.
Nel loro complesso, gli assistenti AI rappresentano un'interfaccia uomo-macchina sempre più centrale nell'ecosistema digitale, in grado di integrare automazione, personalizzazione e accessibilità in un unico sistema intelligente. [26]

4.1.3 IA in salute e fitness

L'intelligenza artificiale assume un ruolo sempre più rilevante nei settori sanitario e del fitness, introducendo soluzioni innovative che migliorano il monitoraggio dello stato di salute e la gestione del benessere individuale. Attraverso l'analisi automatizzata dei dati biometrici e comportamentali, l'IA consente di supportare attività quotidiane come il controllo dei parametri fisiologici, l'analisi delle prestazioni fisiche e la generazione di raccomandazioni personalizzate in ambito sanitario.
Infatti, l'IA si è rivelata particolarmente utile nel monitoraggio glicemico e nella gestione dell'insulina andando a semplificare un'attività che risulta spesso complessa e ripetitiva per i pazienti affetti da diabete.
Per esempio, i sistemi di monitoraggio continuo del glucosio (Continuous Glucose Monitoring) integrano algoritmi di machine learning per analizzare in tempo reale le variazioni della glicemia (vedi Figura 4.2). Questi algoritmi elaborano sia dati correnti sia informazioni storiche per riuscire ad individuare pattern ricorrenti, prevedere possibili fluttuazioni dei livelli di glucosio e, sulla base di queste previsioni, suggerire dosaggi di insulina più accurati e tempestivi.
Questo approccio contribuisce a mantenere i livelli glicemici entro intervalli sicuri, riducendo il rischio di iperglicemia e ipoglicemia e migliorando così la qualità della vita del paziente.

L'introduzione di sistemi di valutazione del corpo basati sull'intelligenza artificiale in ambito fitness consente la creazione di modelli di allenamento altamente personalizzati grazie all'analisi continua di dati biomeccanici, fisiologici e prestazionali che permettono di adattare gli esercizi in maniera dinamica alle reali condizioni e ai progressi maturati.
Il monitoraggio costante dei parametri di movimento, della postura e dei carichi applicati rappresenta inoltre un elemento chiave nella prevenzione degli infortuni: l'analisi automatizzata del movimento consente di individuare precocemente squilibri muscolari, compensazioni scorrette o situazioni di sovraccarico, riducendo il rischio di lesioni e favorendo un'esecuzione più sicura degli esercizi.
Dal punto di vista dell'efficienza, questi sistemi permettono inoltre un'ottimizzazione significativa dei tempi di allenamento: sessioni più mirate e basate su dati consentono di eliminare esercizi ridondanti o poco efficaci.
Infine, l'accesso a dati comparativi rappresenta un valore aggiunto poiché il

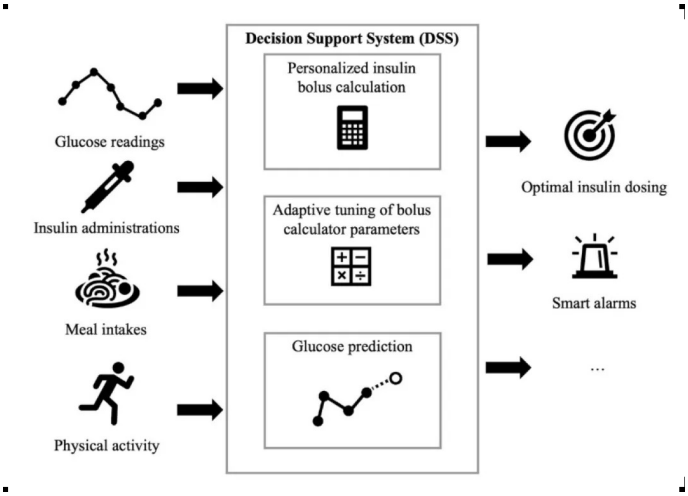

Figura 4.2: Processo di gestione del diabete tramite IA. [27]

confronto in tempo reale delle proprie prestazioni con standard di riferimento consente una valutazione oggettiva facilitando il miglioramento ed il monitoraggio dei risultati.

Nonostante i benefici, l'integrazione tra intelligenza artificiale e fitness o salute introduce alcune importanti sfide. In primo luogo, emerge la necessità di una gestione rigorosa della privacy e della sicurezza dei dati personali: le informazioni raccolte da sensori biometrici e sistemi di visione artificiale rientrano nella categoria dei dati sensibili per cui devono essere protetti con tecniche di cifratura avanzata e politiche di utilizzo trasparenti e conformi alle normative vigenti.
In secondo luogo, è fondamentale mantenere un corretto equilibrio tra l'affidamento ai sistemi automatizzati ed il rispetto della variabilità fisiologica individuale. Non tutti i parametri biologici e biomeccanici rispondono in modo uniforme ai modelli predittivi e un'eccessiva standardizzazione dei protocolli di allenamento potrebbe ridurre l'efficacia. Per questo motivo, i sistemi di IA dovrebbero essere considerati strumenti di supporto decisionale, integrati con la supervisione di professionisti qualificati. [27]

4.1.4 IA nella navigazione

Le navi autonome, note anche come Maritime Autonomous Surface Ships, impiegano un insieme di tecnologie avanzate tra cui intelligenza artificiale, machine learning e Internet of Things (IoT) per svolgere operazioni di navigazione e gestione senza intervento umano diretto.
Si tratta di unità equipaggiate con sensori ad alta precisione (radar, sonar, GPS e telecamere) ed infrastrutture che permettono la raccolta continua di dati sull'ambiente marittimo e sullo stato della nave. Gli algoritmi di AI elaborano tali informazioni in tempo reale per supportare il processo decisionale, ottimizzare le rotte, garantire la sicurezza della navigazione e rispondere ad eventi imprevisti.
Di conseguenza, l'adozione di sistemi autonomi nel settore marittimo permette una riduzione dei costi operativi legata anche alla diminuzione dell'equipaggio ed un miglioramento generale della sicurezza grazie ad un minore errore umano.

Infatti, l'intelligenza artificiale è oggi un elemento chiave nei moderni sistemi di bordo, con applicazioni in diversi ambiti. Nel campo della manutenzione predittiva, grazie all'analisi dei dati provenienti dai sensori sui macchinari, i modelli di machine learning sono in grado di identificare anomalie e prevedere guasti prima che si manifestino, riducendo così i tempi di fermo ed i costi di manutenzione. L'AI viene inoltre impiegata per l'ottimizzazione delle rotte, integrando informazioni su condizioni meteorologiche, correnti oceaniche, traffico navale e vincoli operativi, calcolando percorsi che minimizzano consumo di carburante, tempi di percorrenza ed emissioni. In termini di sicurezza, sistemi intelligenti di monitoraggio continuo analizzano l'ambiente circostante, rilevando potenziali rischi e attivando avvisi o contromisure automatiche per la prevenzione degli incidenti. Infine, l'AI supporta il decision-making autonomo consentendo alle navi di reagire in tempo reale a condizioni dinamiche e imprevedibili.
La piena autonomia operativa rimane un obiettivo futuro, però le tecnologie attuali forniscono già avanzati sistemi di supporto alle decisioni aumentando l'efficacia dell'intervento umano.[28]

4.2 Efficienza on-device

L'on-device AI rappresenta un paradigma dell'intelligenza artificiale in cui l'elaborazione dei dati avviene direttamente sul dispositivo finale, senza il ricorso a infrastrutture di cloud computing. Questo approccio introduce vantaggi rilevanti rispetto alle soluzioni cloud-centriche, in particolare in termini di privacy, latenza e personalizzazione. Il mantenimento dei dati in locale riduce infatti l'esposizione delle informazioni sensibili a rischi legati alla trasmissione e all'archiviazione remota, aspetto fondamentale in contesti che trattano dati personali, biometrici o sanitari.
Inoltre, l'elaborazione locale elimina anche i ritardi dovuti alla comunicazione con server remoti, consentendo tempi di risposta significativamente inferiori e rendendo l'on-device AI particolarmente adatta ad applicazioni che richiedono decisioni in tempo reale, come il riconoscimento vocale, l'analisi delle immagini o il controllo di sistemi autonomi.
Un ulteriore elemento distintivo è la capacità di offrire un elevato grado di personalizzazione: operando su dispositivi personali, i modelli di intelligenza artificiale possono adattarsi ai comportamenti specifici dell'utente e all'utilizzo consentendo di generare risposte più accurate e coerenti con le esigenze individuali senza la necessità di condividere dati con infrastrutture esterne.

Il concetto di on-device AI è strettamente connesso a quello di machine learning e di edge computing, che prevede l'elaborazione dei dati il più vicino possibile alla loro sorgente. In questo senso, l'on-device AI può essere considerata declinazione portata al parossismo dell'edge computing, poiché trasferisce l'intero processo decisionale sul dispositivo stesso. Tale scelta architetturale incrementa non solo la velocità e l'affidabilità del sistema, ma anche la sua capacità di operare in modo autonomo in assenza di una connessione continua alla rete.
Dal punto di vista metodologico invece, il nucleo fondamentale dell'on-device AI è costituito dal machine learning che fornisce ai modelli la capacità di apprendere

dai dati e di adattarsi nel tempo. Tuttavia, le limitazioni hardware tipiche dei dispositivi edge, in termini di potenza di calcolo, memoria ed energia disponibile, impongono l'adozione di modelli computazionalmente efficienti. Tra le soluzioni più diffuse rientrano le reti neurali leggere, progettate per svolgere compiti specifici con un numero ridotto di parametri e i modelli basati su alberi decisionali che offrono rapidità di inferenza e un'elevata interpretabilità con un costo computazionale contenuto.

Le applicazioni dell'intelligenza artificiale on-device sono già ampiamente diffuse in numerosi settori. Nei dispositivi mobili, come smartphone e tablet, l'esecuzione locale dei modelli di AI abilita funzionalità avanzate di riconoscimento vocale, elaborazione delle immagini e suggerimenti personalizzati, migliorando al contempo la tutela della privacy. Nei dispositivi indossabili invece, l'on-device AI consente il monitoraggio continuo di parametri fisiologici e comportamentali, fornendo feedback immediati sullo stato di salute e sull'attività fisica. In ambito smart home, l'elaborazione locale permette a videocamere e termostati intelligenti di reagire rapidamente agli eventi, ottimizzando sicurezza ed efficienza energetica. Nel settore automobilistico, infine, l'on-device AI riveste un ruolo centrale nei sistemi di assistenza alla guida e nelle applicazioni di guida autonoma, dove l'elaborazione in tempo reale è essenziale per garantire sicurezza e affidabilità.
Dunque, l'on-device AI si configura come una componente fondamentale nello sviluppo di sistemi intelligenti capaci di operare in modo autonomo, efficiente e sicuro. Riducendo la dipendenza dal cloud e valorizzando l'elaborazione locale, questo approccio apre la strada a nuove applicazioni in grado di rispondere alle crescenti esigenze di velocità, personalizzazione e protezione dei dati. [29]

4.3 Utilizzo collettivo dei servizi IA

4.3.1 Da un punto di vista economico

Anche il Joint Research Centre dell'Unione Europea ([30]) riconosce le tecnologie digitali come componenti strategiche per la transizione verso un sistema energetico più sostenibile in grado di migliorare l'efficienza complessiva del sistema elettrico.
Il mercato dell'elettricità costituisce un elemento centrale del sistema elettrico poiché rappresenta il luogo in cui produttori, consumatori ed operatori di rete interagiscono con l'obiettivo di garantire in tempo reale l'equilibrio tra domanda e offerta. Un funzionamento inefficiente o distorto di questo mercato può generare conseguenze rilevanti, come forti oscillazioni dei prezzi dell'energia o addirittura interruzioni del servizio e blackout con impatti economici e sociali significativi.
Per capire come l'integrazione dell'IA nel mercato elettrico influenzi la società, si adotta il quadro dei valori pubblici, approccio che permette di valutare come le tecnologie digitali possono rafforzare o, al contrario, mettere sotto pressione il funzionamento dei vari sistemi.
Il quadro distingue nove valori pubblici, ossia nove categorie, per lo studio degli impatti sociali ed etici associati all'adozione di mercati elettrici basati su AI.
Questi valori sono costruiti attraverso processi democratici e incorporati nelle istituzioni che orientano l'azione collettiva verso ciò che è ritenuto giusto o sbagliato

per la società, dove, nel contesto dei sistemi elettrici, vari studi hanno riportato essere la giustizia, il controllo, la fiducia, la sicurezza e la stabilità del sistema.
Inoltre, la ricerca sull'adozione delle tecnologie dell'informazione e della comunicazione, ha sottolineato ulteriori valori rilevanti tra cui la tutela della privacy, l'equilibrio dei rapporti di potere, l'equità, l'uguaglianza e l'autonomia degli attori coinvolti.
Alla luce di queste considerazioni, l'analisi dell'impatto sociale dei mercati elettrici basati su AI funde da riferimento per valutare in che misura l'intelligenza artificiale possa contribuire ad un'evoluzione positiva del sistema energetico.

Grazie alla capacità di elaborare grandi quantità di dati eterogenei in tempi ridotti, i sistemi di AI permettono di migliorare significativamente l'accuratezza delle previsioni. Si tratta di un aspetto fondamentale per la stabilità del sistema elettrico poiché previsioni più precise consentono di anticipare squilibri tra domanda e offerta.
Un altro ambito di applicazione riguarda il funzionamento del mercato elettrico: l'utilizzo di algoritmi di AI consente di automatizzare e accelerare i processi decisionali. Dal punto di vista fisico, ciò aumenta la flessibilità del sistema, permettendo una risposta più rapida alle variazioni di carico e di generazione. La riduzione delle scale temporali di intervento è un fattore determinante per l'integrazione efficiente delle energie rinnovabili e per il mantenimento della frequenza e della stabilità della rete.
L'AI può inoltre essere impiegata nella gestione diretta degli asset energetici, l'algoritmo infatti può intervenire direttamente sul comportamento fisico dei dispositivi, modulando produzione e consumo in funzione delle condizioni di rete e dei segnali di mercato. Una maggiore precisione nelle previsioni e una gestione più rapida del bilanciamento contribuiscono a ridurre l'uso di riserve fossili e a contenere i costi di produzione.
Dal punto di vista energetico invece, si traduce in un aumento dell'efficienza globale del sistema e in una migliore utilizzazione delle risorse rinnovabili disponibili.
Tuttavia, sono presenti anche criticità rilevanti. L'aumento della complessità del sistema introduce nuovi rischi legati alla sicurezza informatica, alla perdita di controllo umano e alla difficoltà di individuare e correggere errori in tempo reale. Un ulteriore elemento di incertezza riguarda la distribuzione del potere e delle risorse all'interno del sistema. Sebbene l'AI possa abbassare le barriere di accesso al mercato per piccoli produttori, nella pratica i benefici maggiori si concentrano su chi è dotato di maggiori capacità computazionali e finanziarie portando ad una crescente centralizzazione del controllo, in contrasto con l'idea di un sistema elettrico realmente distribuito. [31]

4.3.2 Da un punto di vista sociale

Nonostante sembri naturale usare espressioni come "per favore" e "grazie" con ChatGPT, simili gesti di cortesia hanno in realtà un costo non trascurabile in termini economici ed energetici. Il CEO di OpenAI, Sam Altman, ha evidenziato come questi scambi educati abbiano un impatto concreto sulle risorse aziendali: ogni parola inviata a un modello di AI richiede potenza di calcolo per essere elaborata e per generare una risposta che si traduce in un consumo energetico complessivo molto

superiore a quello di una normale ricerca su Google.
Uno studio recente ha rilevato che il 69% dei giovani della Generazione Z usa "per favore" e "grazie" quando interagisce con ChatGPT, il che amplifica il consumo complessivo di risorse. Secondo Microsoft, nel report sulla sostenibilità del 2022, mantenere lo status quo potrebbe compromettere la disponibilità di acqua dolce per le generazioni future. Anche Google avverte che la domanda globale di acqua dolce potrebbe superare l'offerta del 40% entro il 2030, sottolineando l'urgenza di ridurre gli sprechi legati all'energia.
Dal punto di vista sociale, invece, ci sono riflessioni contrastanti sull'educazione verso i chatbot. Alcuni temono che la tolleranza di assistenti vocali verso la maleducazione possa trasmettere cattive abitudini, mentre altri ritengono che trattare le macchine con durezza possa aiutare a distinguere chiaramente gli oggetti intelligenti dagli esseri umani.
Altman, tuttavia, considera il "costo" dell'educazione verso l'AI un investimento ragionevole: rendere i prodotti il più simili possibile agli esseri umani comporta l'adozione di norme sociali, migliorando l'interazione complessiva con i sistemi di intelligenza artificiale. [32]

4.4 Rebound Effect

Il Rebound Effect digitale descrive il fenomeno per cui i miglioramenti di efficienza nelle tecnologie possono tradursi in incrementi inaspettati del consumo complessivo, riducendo o addirittura annullando i risparmi di risorse inizialmente previsti. Si tratta dunque della risposta macroeconomica e comportamentale ai miglioramenti di efficienza tecnologica, risposta che porta però ad una parziale o totale erosione dei benefici ambientali attesi. In alcuni casi, tali dinamiche possono persino determinare un aumento netto del consumo di risorse e dell'impatto ambientale su scala locale o globale.
È pertanto un fenomeno di natura profondamente sistemica: il progresso tecnologico, pur migliorando l'efficienza di singoli processi, non garantisce automaticamente il raggiungimento degli obiettivi di sostenibilità, poiché i cambiamenti indotti nei comportamenti e nelle strutture socio-economiche possono compensare o annullare i benefici iniziali.

L'effetto rebound digitale non si manifesta in modo uniforme, ma assume forme differenti con implicazioni specifiche per la sostenibilità.

- Il rebound diretto rappresenta la modalità più immediata e intuitiva: si verifica all'interno dello stesso servizio in cui vengono ottenuti i guadagni di efficienza. Un esempio emblematico è l'aumento dell'efficienza energetica dei data center che, riducendo il costo per unità di calcolo o di archiviazione, può indurre una crescita della domanda di servizi digitali compensando in parte o del tutto i risparmi energetici iniziali.

- Il rebound indiretto emerge invece da risposte più ampie del sistema economico. I risparmi di costo generati dall'efficienza delle tecnologie digitali possono

propagarsi influenzando prezzi e modelli di consumo. Un caso rappresentativo è quello dell'e-commerce: l'efficienza delle piattaforme digitali e dei sistemi logistici riduce i costi di acquisto e di transazione, incentivando un aumento dei consumi complessivi. Parte di questa crescita può riguardare settori ad alta intensità di risorse, come il trasporto delle merci e la produzione industriale, andando ad amplificare l'impatto ambientale complessivo. Questo tipo di rebound è meno immediato da osservare, ma comunque molto rilevante.

- Il rebound trasformazionale riguarda cambiamenti strutturali nei modelli di consumo, nell'organizzazione economica e negli stili di vita indotti dalla diffusione delle tecnologie digitali. La crescita della gig economy, abilitata dalle piattaforme digitali, ne costituisce un esempio significativo: sebbene le singole transazioni possano risultare efficienti, la trasformazione complessiva del sistema può comportare un aumento del consumo di risorse, ad esempio attraverso una maggiore domanda di trasporti, consegne a domicilio e servizi on-demand. Questo tipo di rebound opera su orizzonti temporali lunghi e contribuisce a ridefinire norme sociali, abitudini di consumo e fabbisogni energetici.

L'intensità dell'effetto rebound digitale non è costante, ma dipende da diversi fattori chiave. Un ruolo centrale è svolto dall'elasticità della domanda rispetto al prezzo: se la domanda di un servizio digitale è molto elastica anche piccole riduzioni di costo possono generare aumenti significativi dei consumi, amplificando l'effetto rebound. Al contrario, in presenza di una domanda rigida, l'impatto tende a essere più contenuto.
Un ulteriore elemento rilevante è il livello di saturazione del servizio: quando la domanda è prossima alla saturazione, come nel caso dell'accesso di base a Internet in molti Paesi ad alto reddito, la crescita dei consumi è limitata, riducendo il potenziale rebound associato a ulteriori miglioramenti di efficienza.
I fattori comportamentali rivestono anch'essi un ruolo cruciale: la percezione dell'efficienza da parte degli utenti influenza le modalità di utilizzo. Le tecnologie considerate "più efficienti" o "meno impattanti" possono essere usate con maggiore frequenza o con minore attenzione ai consumi, contribuendo ad amplificare l'effetto rebound.
Infine, il contesto normativo e politico può attenuare o accentuare l'effetto rebound digitale: in assenza di regolamentazioni adeguate, i guadagni di efficienza rischiano di tradursi in un aumento incontrollato dei consumi. Al contrario, strumenti come la tariffazione del carbonio, standard energetici stringenti e politiche di internalizzazione dei costi ambientali possono contribuire a contenere il rebound, mantenendo elevato il costo del consumo e preservando i benefici ambientali dell'innovazione tecnologica.

Le radici dell'effetto rebound digitale affondano in osservazioni empiriche e in quadri teorici sviluppati nell'economia, nella sociologia e nelle scienze ambientali. Queste dinamiche emergono con chiarezza nell'intersezione tra tecnologie digitali, consumo energetico e lavoro da remoto. Infatti, il telelavoro viene spesso presentato come una pratica sostenibile grazie alla riduzione degli spostamenti quotidiani casa-lavoro. In effetti, le prime analisi suggerivano risparmi energetici rilevanti legati alla diminuzione del traffico pendolare. Tuttavia, studi successivi, basati su

dati più dettagliati e su approcci metodologici più completi, hanno evidenziato la possibile comparsa di un effetto rebound digitale.
Uno dei primi canali attraverso cui questo effetto si manifesta è l'aumento dei consumi energetici domestici: lavorare da casa implica un maggiore utilizzo di energia per riscaldamento o raffreddamento, illuminazione e alimentazione delle postazioni di lavoro. Lo spostamento della domanda energetica da edifici centralizzati verso abitazioni private può così tradursi in un uso complessivamente meno efficiente dell'energia, soprattutto se le abitazioni hanno prestazioni energetiche inferiori rispetto agli uffici moderni, incrementando quindi i consumi nel settore edilizio e attenuando o annullando i risparmi ottenuti dalla riduzione dei trasporti.
Inoltre, il lavoro da remoto può innescare una riconfigurazione dei modelli di consumo. Infatti, il tempo e le risorse economiche risparmiate grazie all'eliminazione degli spostamenti quotidiani possono essere reindirizzati verso altre attività ad alta intensità energetica. Un esempio è l'aumento degli acquisti online, con le emissioni associate ai processi logistici ed alle consegne, oppure una maggiore propensione ai viaggi per svago resa possibile dalla maggiore flessibilità lavorativa. In questo modo, i benefici ambientali iniziali vengono progressivamente compensati da nuove fonti di consumo ed emissioni.
Un ulteriore elemento critico da tenere in considerazione riguarda le infrastrutture digitali necessarie a supportare il lavoro a distanza. L'espansione dei data center, delle reti di telecomunicazione e soprattutto della diffusione di dispositivi elettronici personali comporta un consumo crescente di energia e di risorse materiali. L'impronta ambientale dell'infrastruttura digitale, spesso invisibile all'utente, diventa così una componente centrale dell'analisi che può contribuire all'effetto rebound.
Infine, il lavoro da remoto può favorire cambiamenti più profondi negli stili di vita e nelle scelte residenziali. La minore necessità di recarsi quotidianamente in ufficio può incentivare una maggiore dispersione urbana, con persone che scelgono di vivere più lontano dai centri cittadini. Questa tendenza può aumentare la dipendenza dall'automobile per altre attività quotidiane e alimentare così fenomeni di espansione urbana, con ulteriori conseguenze ambientali: la flessibilità associata al lavoro da remoto può tradursi, in modo non intenzionale, in configurazioni di vita meno sostenibili.

L'insieme di queste dinamiche mostra come l'effetto rebound digitale non si esaurisca in semplici variazioni dirette dei consumi energetici, ma si configuri come una cascata di effetti interconnessi che attraversano diversi settori dell'economia e della società. Il nodo centrale risiede appunto nel carattere sistemico di tali interazioni e nella difficoltà di anticiparle e governarle.
La mitigazione dell'effetto rebound digitale può quindi consistere nell'adozione di valutazioni del ciclo di vita che vadano oltre l'efficienza immediata e considerino l'intero impatto ambientale delle tecnologie digitali, includendo l'energia incorporata nei dispositivi, il consumo delle infrastrutture e le fasi di fine vita. Accanto a ciò, è essenziale integrare strumenti di economia comportamentale nella progettazione dei sistemi digitali e delle politiche pubbliche orientando le scelte degli utenti verso modelli di consumo più sostenibili.
Un ulteriore passo riguarda l'integrazione dell'effetto rebound nelle politiche pubbliche. Strumenti come il carbon pricing, standard di efficienza energetica per le

infrastrutture digitali e politiche ispirate ai principi dell'economia circolare possono aiutare a contenere gli effetti indesiderati dei miglioramenti di efficienza.
Infine, affrontare l'effetto rebound digitale richiede una collaborazione trasversale tra sviluppatori tecnologici, politici, imprese e comunità scientifica: solo attraverso una pianificazione integrata e interdisciplinare è possibile riconoscere la complessità del fenomeno, anticiparne gli impatti e governare le trasformazioni digitali in modo coerente con gli obiettivi di sostenibilità ambientale. [33]

Capitolo 5

Strategie per la riduzione dell'impatto energetico

5.1 Ottimizzazione software e hardware

Negli ultimi decenni, l'incremento delle prestazioni dei processori è di gran lunga cresciuto rispetto ai progressi nelle tecnologie delle batterie, determinando così una riduzione dell'autonomia dei dispositivi mobili, passata da diversi giorni a poche ore di utilizzo. In aggiunta, le applicazioni mobili di nuova generazione richiedono capacità di calcolo sempre più elevate ed un'elaborazione grafica intensiva con vincoli temporali di esecuzione stringenti.
Oggi, questi requisiti vengono soddisfatti principalmente attraverso architetture many-core ad alta frequenza in grado di garantire elevate velocità di elaborazione, ma a discapito di un importante aumento del consumo energetico. Di conseguenza, è necessario adottare approcci di gestione delle risorse innovativi per riuscire a contenere i consumi e, allo stesso tempo, garantire prestazioni elevate.

In questo contesto, i sistemi embedded many-core eterogenei e riconfigurabili rappresentano una possibile soluzione, poiché consentono di migliorare l'efficienza energetica senza però sacrificare le prestazioni. Ciò è possibile grazie ad una combinazione di schedulazione intelligente dei task e adattabilità dell'hardware: i core di dimensioni ridotte e a basso consumo vengono impiegati per l'elaborazione di compiti poco onerosi dal punto di vista computazionale, mentre i core più grandi e performanti vengono attivati solo quando sono richieste elevate prestazioni, accettando un consumo energetico superiore.

5.1.1 Tecniche di efficienza energetica nelle architetture multi-core statiche e omogenee

Oltre alle soluzioni a livello architetturale, sono stati proposti numerosi approcci per ridurre il consumo energetico direttamente all'interno del core. Tra questi, una delle tecniche più diffuse è il Dynamic Voltage and Frequency Scaling (DVFS), che permette di bilanciare dinamicamente consumo energetico e prestazioni tramite la regolazione della tensione di alimentazione e della frequenza di clock. Il principio di

base del DVFS consiste nell'adattare il livello di prestazioni del processore alle reali esigenze dell'applicazione in esecuzione: quando è prioritario rispettare requisiti prestazionali stringenti, vengono selezionati livelli elevati di tensione e frequenza; al contrario, quando la richiesta di calcolo è ridotta si adottano livelli più bassi, in una modalità comunemente nota come CPU throttling.
Dal punto di vista implementativo, l'architettura del DVFS è progettata in modo da consentire al sistema operativo di selezionare dinamicamente la coppia tensione–frequenza desiderata attraverso la scrittura di registri di controllo. Per ciascun livello di prestazione richiesto, il sistema operativo configura il processore nella modalità a minimo consumo energetico compatibile con i vincoli temporali dell'applicazione.
La riduzione della tensione di alimentazione e della frequenza di clock comporta un rallentamento dell'esecuzione delle istruzioni, ma consente un funzionamento energeticamente più efficiente.
Nei circuiti logici, l'aumento dei tempi di salita e discesa dei segnali, insieme all'allungamento del periodo di clock, implica infatti che i vincoli prestazionali debbano essere opportunamente allentati. Nella pratica, le architetture moderne integrano una combinazione di tecniche di risparmio energetico, tra cui DVFS, gestione avanzata dei core e meccanismi di spegnimento selettivo, al fine di massimizzare l'efficienza complessiva, prolungare la durata della batteria e garantire il rispetto dei requisiti prestazionali, soprattutto nei dispositivi mobili. Questa tecnica può essere estesa anche a sistemi multi-core omogenei per emulare un comportamento eterogeneo, regolando indipendentemente tensione e frequenza di ciascun core di modo che, core con architettura identica possono operare con diverse caratteristiche di ritardo e consumo, adattandosi meglio ai requisiti dei task assegnati. Studi sperimentali mostrano che l'integrazione di regolatori di tensione on-chip ad alta velocità che commutano la tensione su scala di nanosecondi, implicano un significativo risparmio energetico.
Ulteriori benefici emergono combinando il DVFS con la migrazione dei thread, consentendo di allocare dinamicamente i carichi di lavoro sui core più appropriati dal punto di vista energetico.
Tuttavia, molte tecniche DVFS tradizionali assumono un singolo hardware per core, risultando meno efficaci. Per superare questo limite, approcci avanzati come il thread shuffling, raggruppano thread con requisiti simili e, coordinando DVFS e migrazione in modo consapevole, consentono di ottenere riduzioni del consumo energetico fino al 56%, senza però penalizzare le prestazioni, evidenziando così il potenziale di una gestione energetica ed integrata nelle architetture multi-core.

In aggiunta al DVFS, l'efficienza energetica dei processori moderni è ottenuta attraverso un insieme di tecniche microarchitetturali complementari, applicate sia in fase di progettazione sia a runtime. Ad esempio, il clock gating, applicato in fase di progettazione, interviene sulla potenza dinamica introducendo logica di controllo tra la sorgente del clock ed i circuiti sincroni. Se inserito nel contesto del Dynamic Power Management, consente di ridurre consumi superflui durante l'esecuzione evitando commutazioni non necessarie ma senza andare a compromettere le prestazioni.
Un'ulteriore riduzione dei consumi è ottenuta tramite i domini di potenza, ossia

regioni di un processore controllate da una singola alimentazione in maniera indipendente, che possono essere completamente spente per minimizzare il consumo energetico senza però rimuovere totalmente l'alimentazione dal sistema. Nelle architetture multi-core, l'adozione di domini di potenza per singolo core permette di spegnere i core inattivi e, allo stesso tempo, di far operare quelli attivi a livelli prestazionali più elevati, riducendo comunque la potenza complessiva del chip. La combinazione di power gating e clock gating amplifica ulteriormente i benefici energetici.

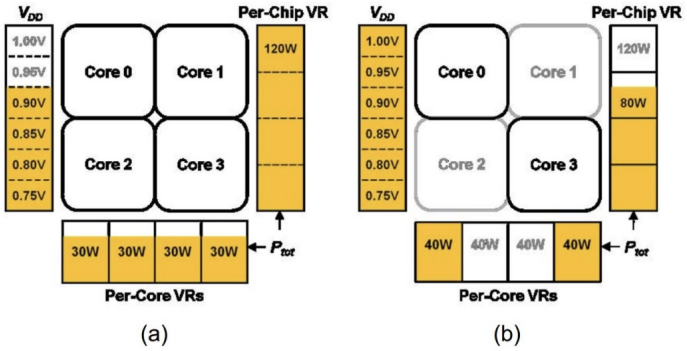

Figura 5.1: Domini di alimentazione per singolo core che possono ridurre il consumo energetico e garantire livelli di prestazioni più elevati. [34]

In Figura 5.1 è mostrato un meccanismo che consente di ridurre il consumo complessivo di potenza del chip mantenendo al contempo le prestazioni per core. Nel caso (a) è illustrato un dominio di potenza unico a livello di chip, con tutti i core attivi allo stesso livello. Al contrario, nel caso (b) è mostrato un dominio di potenza per singolo core, che permette di spegnere i core non necessari e di consentire ai core attivi di operare ad un livello prestazionale più elevato, riducendo la potenza totale del chip.

Un'altra importante tecnica più orientata alla microarchitettura è il bilanciamento della pipeline (Pipeline Balancing): modificando dinamicamente i vincoli di ritardo degli stadi della pipeline in funzione dei requisiti prestazionali, è possibile migliorare l'efficienza energetica in modo analogo al DVFS. L'approccio delle Power Balanced Pipelines mira a ridurre le disparità di consumo tra gli stadi, assegnando ritardi differenziati pur garantendo un determinato livello di prestazioni. Il bilanciamento può essere statico, se eseguito in fase di progettazione, o dinamico, se si adatta al carico di lavoro riducendo il consumo energetico fino al 46%.

Infine, è necessario sottolineare l'importanza del ruolo svolto dalla co-progettazione di core, cache e interconnessioni. Cache di grandi dimensioni possono ostacolare la scalabilità e aumentare il costo energetico dell'interconnessione, mentre interconnessioni estese e non ottimizzate possono risultare energivore. Un equilibrio accurato è quindi essenziale per garantire alte prestazioni con un consumo energetico contenuto. [35]

5.1.2 Tecniche di efficienza energetica nelle architetture multi-core eterogenee

Un'architettura multi-core eterogenea è composta da core con dimensioni e livelli di complessità differenti, progettati per operare in maniera complementare dal punto di vista delle prestazioni e dell'efficienza energetica. I diversi core rappresentano specifici punti nello spazio potenza–prestazioni; di conseguenza, è possibile ottenere benefici energetici significativi allocando dinamicamente l'esecuzione delle applicazioni sul core più adeguato alle richieste del carico di lavoro.
Per fare questo, è necessario implementare un meccanismo di commutazione dei task che sia capace di bilanciare i requisiti prestazionali con il mantenimento dell'efficienza energetica complessiva del sistema. Un esempio emblematico di architettura eterogenea è rappresentato dalla combinazione di un processore compatto ed energeticamente efficiente, come il Cortex-A7, con un processore ad alte prestazioni, come il Cortex-A15. Si tratta di una configurazione concepita per tener conto dei pattern di utilizzo dinamici tipici degli smartphone moderni, nei quali brevi intervalli di elaborazione intensiva si alternano a periodi più lunghi con una bassa intensità computazionale.
In questo modo, i compiti a ridotto carico computazionale come la messaggistica o l'elaborazione audio, possono essere eseguiti sul core A7, consentendo un risparmio energetico significativo ed una maggiore durata della batteria. Al contrario, quando si presentano fasi di carico elevato, viene attivato il core A15 per soddisfare vincoli prestazionali più stringenti.
I vantaggi energetici che derivano dall'eterogeneità possono però essere pienamente sfruttati solo attraverso una corretta assegnazione dei task o delle applicazioni ai core disponibili, in modo da massimizzare l'efficienza energetica ma garantendo il rispetto delle scadenze prestazionali.
L'allocazione dei task nei sistemi multi-core può essere ulteriormente ottimizzata considerando le diverse fasi di esecuzione dei programmi. Infatti, le applicazioni attraversano fasi caratterizzate da differenti esigenze di risorse computazionali: riconoscere e sfruttare tali variazioni permette di migliorare simultaneamente prestazioni ed efficienza energetica.
Approcci capaci di adattare in tempo reale l'assegnazione dei programmi ai core in risposta alle variazioni comportamentali delle applicazioni, mostrano importanti riduzioni energetiche rispetto a schemi statici, evidenziando come la dinamicità del carico di lavoro rappresenti un elemento chiave per il miglioramento dell'efficienza energetica complessiva. [35]

5.1.3 Tecniche di efficienza energetica nelle architetture riconfigurabili

La riconfigurabilità è una proprietà necessaria per incrementare l'efficienza energetica nei processori e nei sistemi su chip poiché si tratta di architetture che introducono adattabilità e flessibilità hardware, permettendo al sistema di modificare dinamicamente la propria struttura in funzione dell'applicazione in esecuzione. A differenza delle architetture eterogenee tradizionali, che combinano core differenti ma statici, le architetture riconfigurabili consentono di ottenere simultaneamente prestazioni elevate e alta efficienza energetica all'interno dello stesso processore.
Infatti, l'attenzione si concentra sulle architetture riconfigurabili a runtime, esclu-

dendo quelle che rimangono statiche durante il funzionamento. Una delle tecniche più rilevanti in questo ambito è la Dynamic Partial Reconfiguration che, come espresso nel nome, permette di riconfigurare parzialmente il circuito durante l'esecuzione del sistema. Solo una porzione dell'architettura viene modificata, il resto continua ad operare normalmente consentendo di adattare l'hardware in tempo reale alle esigenze dell'applicazione, attivando acceleratori hardware specifici solo quando necessario e disattivandoli inseguito per ridurre il consumo energetico. Dunque, è un approccio che ottimizza anche l'uso delle risorse hardware ed incrementa le prestazioni complessive poiché più applicazioni possono essere eseguite in parallelo senza richiedere una riconfigurazione completa dell'intero sistema. Inoltre, la riconfigurazione parziale riduce i tempi ed i costi energetici rendendo il sistema più reattivo ed efficiente.
Tuttavia, l'operazione di riconfigurazione introduce un overhead energetico e temporale che può annullare i vantaggi se gli acceleratori hardware vengono utilizzati per intervalli troppo brevi. La presenza di una regione statica ampia, necessaria per il controllo e la gestione del sistema, comporta inoltre un consumo di potenza di base costante. A ciò si aggiunge che le tipiche interconnessioni programmabili richiedono un numero maggiore di porte logiche, con conseguente aumento del consumo energetico e riduzione delle prestazioni.

Dunque, l'analisi congiunta delle tecniche di assegnazione dinamica dei task, delle architetture eterogenee e delle soluzioni riconfigurabili mette in mostra come l'efficienza energetica nei sistemi multi-core moderni non possa più essere affrontata con approcci statici. Infatti, la chiave risiede proprio nella capacità del sistema di adattarsi continuamente alle variazioni del carico di lavoro, sfruttando flessibilità architetturale, riconfigurazione hardware e strategie di schedulazione intelligenti. [35]

5.2 Algoritmi efficienti

L'efficienza di un algoritmo rappresenta una misura quantitativa della sua capacità di risolvere un problema utilizzando in maniera ottimale le risorse computazionali disponibili, principalmente tempo di esecuzione e memoria. Viene valutata in funzione della dimensione dell'input ed è un concetto centrale nell'informatica applicato in ambiti come la statistica computazionale, la data science ed il machine learning, dove l'elaborazione di grandi volumi di dati è la norma.
Un elemento fondamentale dell'efficienza algoritmica è la complessità temporale, che descrive come il tempo di esecuzione di un algoritmo cresce al variare della dimensione dell'input. Questa crescita consente di prevedere la scalabilità di un algoritmo e di valutare se rimane utilizzabile al crescere della quantità di dati. Accanto alla complessità temporale, la complessità spaziale misura la quantità di memoria richiesta da un algoritmo in funzione dell'input. Include sia lo spazio occupato dai dati di input sia la memoria ausiliaria necessaria durante l'esecuzione. La minimizzazione della complessità spaziale è particolarmente rilevante in contesti con risorse limitate, quali dispositivi mobili e applicazioni edge.

La crescita temporale viene comunemente espressa mediante la notazione O (O grande), strumento matematico che fornisce una stima asintotica del comportamento dell'algoritmo descrivendo un limite superiore alla crescita del consumo di tempo o spazio al crescere dell'input. Classi di complessità comuni includono $O(1)$ per algoritmi a tempo costante, $O(n)$ per tempo lineare, $O(n \log n)$ per algoritmi efficienti di ordinamento, e $O(n^2)$ per algoritmi quadratici. Si tratta dunque di una notazione essenziale per confrontare algoritmi alternativi e comunicare rigorosamente le prestazioni.

Ci sono diversi fattori che influenzano l'efficienza di un algoritmo. Tra questi rientrano la scelta delle strutture dati, la natura dell'algoritmo (iterativo o ricorsivo) ed il tipo di operazioni effettuate.
Dal punto di vista applicativo, l'efficienza algoritmica risulta fondamentale in numerosi ambiti, dall'ottimizzazione dei motori di ricerca all'elaborazione di grandi moli di dati, fino all'addestramento e all'inferenza dei modelli di machine learning. In quest'ultimo caso, algoritmi più efficienti possono ridurre drasticamente i tempi di addestramento e consentire l'elaborazione di flussi di dati in tempo reale, migliorando anche la scalabilità e la sostenibilità computazionale dei sistemi.
La valutazione dell'efficienza di un algoritmo avviene generalmente attraverso una combinazione di analisi teorica ed esperimenti empirici. L'analisi teorica, basata sulla notazione O grande permette di stimare il comportamento asintotico dell'algoritmo indipendentemente dall'hardware utilizzato. I test empirici invece, consentono di misurare il tempo di esecuzione e l'utilizzo della memoria in scenari reali, fornendo una validazione pratica delle prestazioni teoriche. [36]

Nel contesto dell'energia e della sostenibilità, un algoritmo efficiente può essere definito come una procedura computazionale progettata per risolvere un problema minimizzando il consumo di risorse critiche.
Gli algoritmi efficienti contribuiscono alla sostenibilità attraverso diversi meccanismi. In primo luogo, riducono il tempo di elaborazione ed il numero di operazioni necessarie, abbattendo il consumo energetico dei processori. In secondo luogo, ottimizzano l'allocazione delle risorse, limitando movimenti di dati superflui e operazioni ridondanti, che rappresentano una quota rilevante del consumo energetico nei sistemi moderni. Infine, algoritmi meno onerosi dal punto di vista computazionale possono essere eseguiti su hardware meno potente, riducendo l'impronta energetica complessiva dell'infrastruttura di calcolo.
Quando l'efficienza viene analizzata in una prospettiva di sostenibilità energetica, il suo significato si estende oltre i tradizionali indicatori prestazionali. Un algoritmo energeticamente efficiente contribuisce direttamente alla riduzione dell'impatto ambientale dei sistemi computazionali, diminuendo il consumo di energia e, di conseguenza, le emissioni di gas serra associate. Questo aspetto assume particolare rilievo nei sistemi su larga scala, dove anche inefficienze marginali, se replicate milioni di volte, possono tradursi in un consumo energetico significativo.
Il ruolo degli algoritmi efficienti risulta evidente in ambiti ad alta intensità energetica, come le reti elettriche intelligenti. Questi sistemi complessi si basano su processi decisionali algoritmici per la previsione della domanda, il bilanciamento dei carichi, l'individuazione dei guasti e l'integrazione delle fonti rinnovabili. Algoritmi ineff-

ficienti possono generare sovrapproduzione, instradamenti non ottimali o maggiori perdite di trasmissione. Al contrario, algoritmi efficienti consentono una distribuzione ottimizzata dell'energia, una gestione dinamica della domanda ed un utilizzo più efficace delle fonti rinnovabili, riducendo la dipendenza dai combustibili fossili e migliorando la resilienza del sistema energetico. In questo contesto, l'efficienza non riguarda esclusivamente la velocità di calcolo, ma assume una valenza sistemica e ambientale.
La distinzione tra algoritmi efficienti ed inefficienti emerge chiaramente analizzando l'utilizzo delle risorse nelle fasi di computazione, comunicazione e archiviazione. Un esempio emblematico è la compressione dei dati: riducendo la quantità di informazioni da trasmettere o memorizzare, si diminuisce il consumo energetico delle reti di comunicazione e dei sistemi di storage. Su larga scala, questo si traduce in una riduzione significativa del fabbisogno energetico dei data center, inclusi i costi di raffreddamento e gestione operativa.
L'importanza dell'efficienza algoritmica è ulteriormente accentuata dalla diffusione del cloud computing e dall'adozione massiva di tecniche di intelligenza artificiale e machine learning, caratterizzate da carichi computazionali elevati.
L'addestramento e l'inferenza di modelli complessi richiedono grandi quantità di energia, pertanto, lo sviluppo di algoritmi più efficienti è una condizione necessaria per rendere tali tecnologie compatibili con gli obiettivi di sostenibilità ambientale. [37]

5.3 Distillazione

La distillazione della conoscenza (Knowledge Distillation, KD) è una tecnica di machine learning finalizzata al trasferimento delle conoscenze apprese da un modello di grandi dimensioni, modello insegnante, a un modello più compatto, modello studente. È ampiamente utilizzata nel deep learning come metodo di compressione dei modelli e di trasferimento delle capacità predittive, in particolare nel contesto delle reti neurali profonde di grande scala.
L'obiettivo principale della KD è addestrare un modello studente affinché replichi il comportamento di un modello insegnante più complesso. A differenza dell'addestramento convenzionale, che cerca di minimizzare la discrepanza tra le previsioni del modello e le etichette di riferimento del dataset, la KD si concentra sull'allineamento delle predizioni del modello studente con quelle del modello insegnante, sfruttando quest'ultimo come sorgente di informazione più ricca.
Questa tecnica risulta particolarmente rilevante nel contesto attuale, caratterizzato dalla rapida diffusione di modelli di intelligenza artificiale generativa con miliardi di parametri. Tali modelli, pur offrendo prestazioni elevate, risultano spesso inadatti all'impiego pratico a causa dei costi computazionali, dei vincoli di latenza e delle limitazioni hardware. Al contrario, modelli più piccoli garantiscono efficienza e rapidità di inferenza, ma soffrono di una ridotta capacità di generalizzazione e di una minore accuratezza su compiti complessi. La distillazione della conoscenza si pone quindi come una soluzione per combinare i vantaggi di entrambe le categorie.

Il concetto di distillazione è stato introdotto nel 2006 con il lavoro Model Compression, in cui un grande modello d'insieme veniva utilizzato per etichettare

un ampio dataset, successivamente impiegato per addestrare una singola rete neurale compatta. Il modello risultante, pur essendo di alcuni ordini di grandezza più piccolo, mostrava prestazioni comparabili a quelle del modello originale. Successivamente, nel 2015, è stata formalizzata una formulazione più generale della KD, basata sull'idea di separare le fasi di addestramento e di distribuzione: modelli grandi e complessi vengono utilizzati per estrarre la struttura latente dai dati, mentre modelli più piccoli vengono ottimizzati per l'impiego operativo.
Un aspetto centrale della distillazione della conoscenza è la concezione astratta di "conoscenza" non come insieme di parametri appresi, ma come funzione di mappatura dagli input agli output. Infatti, l'obiettivo è addestrare il modello studente a imitare il comportamento funzionale del modello più grande, ovvero la sua capacità di generalizzare a nuovi dati.

Dal punto di vista operativo, la KD introduce una funzione di perdita aggiuntiva, detta distillation loss, che misura la divergenza tra le distribuzioni di output del modello insegnante e di quello studente. Nei modelli più avanzati, la distillazione può includere non solo gli output finali, ma anche rappresentazioni intermedie e sequenze di ragionamento, consentendo allo studente di emulare aspetti più profondi del processo decisionale dell'insegnante.
Le tecniche di distillazione della conoscenza possono essere classificate in base al ruolo del modello insegnante: nella distillazione offline, il modello insegnante è preaddestrato ed i parametri rimangono congelati, è un approccio comune nel caso dei grandi modelli linguistici proprietari; nella distillazione online, invece, insegnante e studente vengono addestrati simultaneamente, consentendo un adattamento dinamico alle variazioni dei dati. Un'ulteriore variante è rappresentata dall'autodistillazione, in cui un singolo modello trasferisce la conoscenza dai livelli più profondi a quelli più superficiali, migliorando l'efficienza dell'inferenza senza però compromettere la capacità di apprendimento durante l'addestramento.
Con l'avvento dei modelli linguistici di grandi dimensioni, la distillazione della conoscenza ha assunto un ruolo strategico poiché consente di trasferire capacità avanzate, come il ragionamento complesso, lo stile linguistico e l'allineamento alle preferenze umane, da modelli di grandi dimensioni a modelli più piccoli. Questo processo permette l'impiego di modelli efficienti su dispositivi edge o in ambienti con risorse limitate, riducendo al contempo costi, latenza ed impatto energetico. [38]

5.4 Modelli SLM

I modelli linguistici di piccole dimensioni (Small Language Models, SLM) sono sistemi di intelligenza artificiale progettati per l'elaborazione, la comprensione e la generazione di linguaggio naturale, caratterizzati da una scala inferiore rispetto ai Large Language Models (LLM). In termini dimensionali, gli SLM includono un numero di parametri che varia da pochi milioni a pochi miliardi, mentre gli LLM possono raggiungere centinaia di miliardi o addirittura trilioni di parametri che comprendono pesi e bias appresi durante la fase di addestramento, determinano il comportamento e le prestazioni del modello.
Recentemente è stato mostrato come LLM e SLM possano essere utilizzati in modo complementare attraverso approcci ibridi. In queste architetture, modelli di piccole

dimensioni operano localmente, mentre modelli di grandi dimensioni vengono interrogati solo quando sono richieste capacità di ragionamento più avanzate. Strategie di routing intelligente consentono di indirizzare dinamicamente le richieste verso il modello più appropriato, ottimizzando l'uso delle risorse.
In realtà. grazie alla loro ridotta complessità, gli SLM risultano più compatti ed efficienti dal punto di vista computazionale: richiedono minori risorse di memoria e di calcolo, rendendoli particolarmente adatti a contesti come dispositivi edge, applicazioni mobili o scenari in cui l'inferenza deve avvenire offline, senza accesso continuo alla rete. In molti casi, gli SLM derivano direttamente da modelli di grandi dimensioni, dai quali ereditano l'architettura di base, tipicamente fondata sui transformer, che rappresentano lo standard di riferimento nell'elaborazione del linguaggio naturale.
Poi, i codificatori trasformano le sequenze di input in rappresentazioni vettoriali che catturano informazioni posizionali e i decodificatori generano la sequenza di output più probabile in base alle rappresentazioni apprese.

La realizzazione di modelli linguistici di piccole dimensioni si basa su tecniche di compressione del modello, volte a ridurre la dimensionalità mantenendo il più possibile l'accuratezza. Tra le principali strategie rientrano il pruning, la quantizzazione, la fattorizzazione a basso rango e la distillazione della conoscenza. Il pruning consiste nell'eliminazione di parametri poco rilevanti o ridondanti, come pesi, neuroni o interi strati della rete, operazione che richiede spesso una successiva fase di fine-tuning per recuperare eventuali perdite di precisione. La quantizzazione riduce la precisione numerica dei parametri, ad esempio passando da rappresentazioni a 32 bit a interi a 8 bit, con un significativo beneficio in termini di carico computazionale e latenza: può essere applicata durante l'addestramento (Quantization-Aware Training) o successivamente (Post-Training Quantization), con un compromesso tra semplicità e accuratezza. La fattorizzazione a basso rango approssima matrici di grandi dimensioni mediante matrici più compatte, riducendo il numero di parametri e di operazioni, ma introducendo una maggiore complessità implementativa. Infine, la distillazione della conoscenza, come riportato nel paragrafo precedente, prevede il trasferimento delle informazioni apprese da un modello "insegnante" di grandi dimensioni a un modello "studente" più piccolo, che viene addestrato a replicarne non solo le predizioni, ma anche il comportamento statistico complessivo colmando il divario tra prestazioni elevate e sostenibilità computazionale.

Nonostante le dimensioni ridotte, gli SLM presentano numerosi vantaggi, dimostrando che la scala non è l'unico fattore determinante per l'efficacia. Infatti, risultano più accessibili consentendo la sperimentazione senza infrastrutture computazionali avanzate; sono più efficienti in termini di addestramento; offrono latenze inferiori grazie al numero ridotto di parametri attivi durante l'inferenza; garantiscono un maggiore controllo sulla privacy potendo quindi essere distribuiti in ambienti privati; presentano un'impronta energetica inferiore, risultando più sostenibili e comportano costi operativi e infrastrutturali significativamente ridotti.
Tuttavia, gli SLM condividono anche alcune criticità tipiche dei modelli di intelligenza artificiale: possono ereditare bias dai modelli di origine, presentare una minore

capacità di generalizzazione su compiti complessi o su domini molto ampi.
Grazie alla possibilità di essere ottimizzati e adattati a domini specifici mediante fine-tuning, gli SLM trovano applicazione in numerosi scenari pratici tra cui i chatbot che rappresentano uno dei casi d'uso più rilevanti: la bassa latenza e l'efficienza computazionale rendono i modelli linguistici di piccole dimensioni ottimali a sistemi di assistenza automatica in cui il modello non si limita a rispondere, ma è in grado di eseguire compiti per l'utente. [39]

5.5 Trend futuri e modelli a basso impatto energetico

Per affrontare l'elevato costo energetico associato ai moderni sistemi di intelligenza artificiale, si sta esplorando lo sviluppo di architetture di AI capaci di operare con consumi dell'ordine di 20 W. Questo valore è paragonabile alla potenza necessaria per alimentare due lampadine LED per un'intera giornata ed è dello stesso ordine di grandezza del consumo energetico quotidiano del cervello umano, rappresentando un obiettivo di efficienza estremamente ambizioso rispetto alle soluzioni di calcolo attualmente impiegate.

5.5.1 Neuromorphic computing

Al centro di queste ricerche si colloca il neuromorphic computing, termine che indica letteralmente sistemi "a forma di cervello o di neuroni". Con esso si fa riferimento sia a circuiti che tentano di riprodurre il comportamento fisico di neuroni e sinapsi biologiche, sia a modelli di calcolo che si ispirano concettualmente ai meccanismi con cui il cervello elabora e memorizza l'informazione (vedi Figura 5.2). A differenza delle architetture convenzionali per l'AI, basate su supercomputer digitali e su elaborazioni binarie di miliardi o trilioni di parametri, i sistemi neuromorfici sfruttano reti elettriche e fotoniche ad altissima efficienza energetica, progettate per emulare direttamente le dinamiche delle reti neurali biologiche. Qui, l'hardware non rappresenta un semplice supporto all'esecuzione di algoritmi, ma diventa parte integrante del processo di inferenza.

Un approccio al calcolo ispirato al cervello consiste nella costruzione di modelli matematici semplificati dei neuroni e delle sinapsi dove i neuroni sono descritti come funzioni non lineari statiche basate su moltiplicazioni scalari, paradigma che, applicato a reti di grandi dimensioni, porta al deep learning. Tuttavia, esistono anche strategie più sofisticate, che introducono dinamiche neuronali e sinaptiche esplicite o che utilizzano impulsi discreti (spike) al posto di segnali continui, avvicinandosi maggiormente al comportamento biologico.

Un'alternativa antitetica ai modelli puramente digitali è rappresentata dagli approcci analogici, che impiegano materiali avanzati in grado di memorizzare un continuo di valori di conduttanza compresi tra 0 e 1. In questi sistemi, le operazioni di moltiplicazione vengono realizzate sfruttando direttamente la legge di Ohm

$$V = RI \tag{5.1}$$

mentre l'accumulazione delle somme parziali avviene tramite la legge di Kirchhoff (legge dei nodi)

$$\sum_{k=1}^{n} I_k = 0 \tag{5.2}$$

Il risultato è una computazione integrata, parallela e con un consumo energetico estremamente ridotto.
Una caratteristica comune a molte architetture neuromorfiche e brain-inspired è la presenza di memoria on-chip, o in-memory computing. Questo approccio rappresenta una rottura concettuale rispetto all'architettura di von Neumann, nella quale memoria e unità di calcolo sono fisicamente separate e i dati devono essere continuamente trasferiti tra le due. Nel cervello umano, al contrario, elaborazione e memorizzazione dell'informazione sono strettamente co-localizzate all'interno delle sinapsi e dei circuiti neurali. Riprodurre questa co-localizzazione nei sistemi artificiali consente di superare il cosiddetto collo di bottiglia di von Neumann, che costituisce uno dei principali limiti energetici e prestazionali dei carichi di lavoro tipici dell'intelligenza artificiale, caratterizzati da un elevato numero di operazioni semplici ma da un'intensa movimentazione dei dati.

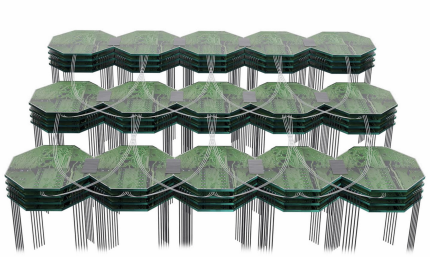

Figura 5.2: Illustrazione di un neuromorphic computer. Ogni wafer ospita un miliardo di sinapsi. I fili bianchi rappresentano connessioni in fibra ottica. [40]

Un contributo fondamentale allo sviluppo del calcolo neuromorfico è stato fornito da Carver Mead, ricercatore presso il California Institute of Technology, che negli anni Novanta dimostrò la possibilità di realizzare dispositivi analogici capaci di riprodurre, a livello fenomenologico, il comportamento di scarica dei neuroni biologici. In questa linea di ricerca si collocano i dispositivi basati su memorie a cambiamento di fase (Phase-Change Memory, PCM), nei quali le unità di calcolo svolgono simultaneamente il ruolo di elaborazione e di memorizzazione dei pesi sinaptici. Le celle PCM, programmate tramite impulsi di corrente che modificano lo stato fisico di materiali vetrosi, consentono di rappresentare i pesi sinaptici come valori di conduttanza analogici, in modo concettualmente simile alla plasticità sinaptica biologica.
In queste architetture, le sinapsi non si limitano a conservare l'informazione, ma partecipano al calcolo. permettendo non solo di superare il collo di bottiglia di von Neumann, ma anche di andare oltre la logica binaria tipica dei transistor digitali, rappresentando valori intermedi con bassi consumi energetici.
Un limite attuale di questi dispositivi analogici risiede però nella fase di addestramento: mentre l'inferenza può essere eseguita in modo efficiente, la modifica diretta

e precisa dei pesi sinaptici durante il training non è ancora sufficientemente controllabile. Per questo motivo, l'addestramento avviene tipicamente su architetture digitali, mentre i pesi appresi vengono successivamente trasferiti nei dispositivi analogici.
Tecnologie affini, come la Resistive Random-Access Memory (RRAM), seguono principi simili consentendo un'elevata parallelizzazione delle operazioni di inferenza. In questo caso, i pesi sinaptici sono codificati nella resistenza di un filamento atomico all'interno di un materiale isolante, la cui conducibilità viene modulata tramite l'applicazione di una tensione.

Anche il flusso dei dati nei chip neuromorfici differisce da quello dei sistemi convenzionali. Nei processori digitali sincroni, il flusso di informazione è governato da un clock globale che sincronizza tutte le operazioni. I sistemi biologici, invece, utilizzano una comunicazione impulsiva ed event-driven: i neuroni trasmettono segnali solo quando necessario, contribuendo così all'elevata efficienza energetica del cervello. Riprodurre un'elaborazione asincrona di questo tipo nei sistemi artificiali potrebbe portare a risparmi energetici significativi, sebbene la realizzazione efficiente di meccanismi spike-based su transistor al silicio presenti ancora diverse sfide tecnologiche.
Nonostante tali difficoltà, i risultati sperimentali più recenti confermano il potenziale del calcolo neuromorfico: test di inferenza su modelli da 3 miliardi di parametri hanno mostrato prestazioni fino a 46,9 volte superiori rispetto alle GPU più efficienti e un'efficienza energetica 72,7 volte maggiore rispetto alle soluzioni digitali a latenza minima. Per queste ragioni, vi è un ampio consenso sul fatto che il calcolo neuromorfico e brain-inspired sia particolarmente adatto ai dispositivi edge, come smartphone e veicoli autonomi, che richiedono inferenza rapida, affidabile ed energeticamente sostenibile su modelli pre-addestrati. [41]

5.5.2 Edge AI

L'Edge AI indica l'implementazione di algoritmi e modelli di intelligenza artificiale direttamente su dispositivi periferici (edge devices), come sensori e dispositivi dell'Internet of Things (IoT). In questo modo, l'elaborazione e l'analisi dei dati avvengono localmente riducendo la dipendenza continua da infrastrutture di cloud computing.
L'Edge AI nasce dall'integrazione tra edge computing e machine learning, consentendo l'esecuzione di modelli di IA direttamente sui dispositivi interconnessi: l'edge computing permette l'archiviazione e la gestione dei dati vicino al punto di acquisizione, mentre gli algoritmi di IA consentono l'estrazione di informazione e la presa di decisioni localmente, anche in assenza di una connessione permanente a Internet. Questo approccio che prevede processi decisionali eseguiti in loco, rende possibile un'elaborazione a bassa latenza, tipicamente dell'ordine dei millisecondi, permettendo risposte e feedback in tempo reale.
Tuttavia, nella maggior parte delle architetture attuali, il cloud rimane necessario per il riaddestramento dei modelli, l'aggregazione dei dati su larga scala e la distribuzione delle versioni aggiornate degli algoritmi di IA.
I sistemi di Edge AI impiegano modelli di machine learning e deep learning tipicamente addestrati in ambienti cloud o data center, dove sono disponibili

grandi volumi di dati e risorse computazionali adeguate. Una volta distribuiti sui dispositivi edge, tali modelli eseguono inferenza localmente. I dati che generano incertezze o errori vengono inviati al cloud per un riaddestramento e i modelli aggiornati vengono successivamente ridistribuiti ai nodi edge, realizzando un ciclo di feedback continuo.

In questo contesto si inserisce il paradigma dell'intelligenza artificiale distribuita (Distributed AI), che estende l'Edge AI introducendo meccanismi di coordinamento, adattamento e monitoraggio di molteplici nodi periferici.
L'IA distribuita consente di automatizzare il ciclo di vita dei dati e dei modelli e coordinare l'esecuzione di compiti e decisioni su una rete di dispositivi eterogenei. Questo approccio permette di scalare le applicazioni di IA su un elevato numero di nodi edge, mantenendo al contempo un elevato grado di autonomia locale.
L'Edge AI abilita inoltre analisi e inferenza in tempo reale anche in assenza di rete continua, aumentando l'affidabilità operativa dei sistemi distribuiti. Tuttavia, la capacità computazionale e di memoria dei dispositivi edge rimane limitata; per questo motivo, viene generalmente integrata con il cloud computing che svolge un ruolo centrale nelle fasi di addestramento dei modelli, aggregazione dei dati e aggiornamento degli algoritmi.

Le differenze principali tra Edge AI e Cloud AI sono:

- Potenza di calcolo: la cloud AI dispone di risorse computazionali e di storage elevate adatte all'addestramento di modelli complessi. L'Edge AI è invece vincolata dalle limitazioni hardware dei dispositivi periferici.
- La latenza rappresenta un fattore critico per molte applicazioni. L'Edge AI riduce drasticamente i tempi di risposta elaborando i dati localmente, mentre la cloud AI introduce ritardi dovuti alla trasmissione dei dati verso server remoti.
- Larghezza di banda: l'elaborazione locale tipica dell'Edge AI riduce la quantità di dati trasmessi in rete. La cloud AI richiede invece un'elevata larghezza di banda per il trasferimento continuo dei dati.
- Sicurezza e privacy: l'Edge AI migliora la protezione dei dati sensibili, che rimangono sul dispositivo o all'interno della rete locale riducendo l'esposizione a rischi associati alla trasmissione verso reti esterne. La cloud AI comporta invece il trasferimento dei dati verso infrastrutture esterne, aumentando i potenziali rischi legati alla sicurezza e alla privacy.

L'Edge AI risulta particolarmente indicata in tutti quei contesti in cui sono richieste previsioni e decisioni in tempo reale come nel controllo automatico, nella robotica e nei sistemi autonomi. Per contro, la cloud AI si riferisce all'implementazione di algoritmi e modelli di intelligenza artificiale su server remoti ad alte prestazioni. Questo approccio offre una capacità di calcolo e di archiviazione significativamente superiore, rendendo possibile l'addestramento e l'utilizzo di modelli di grandi dimensioni e ad alta complessità, a scapito però di una maggiore latenza e di un maggiore consumo di banda.

La rapida espansione dell'Edge AI è dovuta alla crescente diffusione di servizi di edge computing in ambito IoT e dai vantaggi architetturali intrinseci offerti dall'elaborazione locale dell'intelligenza artificiale che consente tempi di risposta dell'ordine dei millisecondi, risultando fondamentale per applicazioni real-time e sistemi cyber-fisici.
Rimangono comunque criticità legate alla protezione fisica e logica dei dispositivi edge, che possono rappresentare punti vulnerabili se non adeguatamente protetti.
Un aspetto chiave dell'Edge AI è inoltre la scalabilità: le architetture ibride edge–cloud consentono l'espansione del sistema mediante l'integrazione di funzionalità edge direttamente nei dispositivi, sia a livello hardware sia software. Questa distribuzione dei carichi di lavoro migliora la tolleranza ai guasti e permette il funzionamento locale anche in caso di interruzioni della connettività verso nodi centrali.
Dal punto di vista economico, l'Edge AI contribuisce alla riduzione dei costi operativi, poiché il cloud viene utilizzato principalmente per l'archiviazione e l'analisi differita dei dati, anziché per l'elaborazione immediata. La distribuzione del carico computazionale riduce l'impiego intensivo di CPU, GPU e memoria nei data center, rendendo l'Edge AI una soluzione più sostenibile rispetto ai modelli di cloud computing puramente centralizzati. Inoltre, l'autonomia decisionale dei dispositivi edge riduce la necessità di supervisione continua, contribuendo a ulteriori risparmi operativi.

L'Edge AI trova applicazione in numerosi settori ad alta intensità di dati e con requisiti stringenti di latenza. In ambito sanitario, consente il monitoraggio continuo dei parametri vitali e l'assistenza in tempo reale nei contesti di emergenza. Nel settore manifatturiero, permette controllo qualità e ottimizzazione dei processi produttivi. Nei sistemi di vendita automatizzata e nelle smart home, migliora l'esperienza utente garantendo risposte rapide e maggiore tutela della privacy. Infine, nei sistemi di sicurezza e sorveglianza, l'elaborazione locale di flussi video permette l'identificazione immediata di eventi critici, superando i limiti di latenza delle soluzioni basate esclusivamente sul cloud.
Altre applicazioni tipiche dell'Edge AI includono veicoli a guida autonoma, dispositivi indossabili, sistemi di videosorveglianza, elettrodomestici intelligenti e, più in generale, sistemi cyber-fisici che richiedono reazioni immediate all'ambiente circostante. In questi contesti, la capacità di elaborare i dati localmente è cruciale per garantire affidabilità, sicurezza e continuità operativa. [42]

5.6 Comparazione tecniche di efficienza energetica nei sistemi di calcolo

Tramite l'analisi delle tecniche per la riduzione dell'impatto energetico nei sistemi computazionali (vedi Tabella 5.1), è stato mostrato come l'efficienza energetica emerga dall'integrazione di strategie su differenti livelli del sistema, dalla microarchitettura hardware fino agli algoritmi e ai modelli di intelligenza artificiale.

Tecnica	Livello di intervento	Risultato	Stato della soluzione
DVFS	Microarchitettura CPU	Riduzione consumo fino al 56%	Ampiamente diffusa
Clock gating	Microarchitettura pipeline	Riduzione della potenza dinamica	Diffusa industrialmente
Power gating	Architettura di sistema	Spegnimento selettivo dei core inattivi	Diffusa industrialmente
Pipeline balancing	Microarchitettura	Riduzione consumo fino al 46%	Adottata in alcune architetture
Architetture multicore eterogenee	Architettura di sistema	Allocazione dinamica dei task su core diversi	Diffusione industriale
Architetture riconfigurabili	Architettura hardware	Adattamento dinamico dell'hardware al carico	Ricerca e applicazioni specialistiche
Algoritmi efficienti	Livello software	Riduzione delle operazioni computazionali	Diffusione generale
Knowledge Distillation	Modelli di IA	Compressione del modello insegnante in uno studente	Diffusione industriale
Small Language Models (SLM)	Modelli di IA	Riduzione della complessità computazionale	Diffusione crescente
Edge AI	Architettura distribuita	Inferenza locale con latenza di millisecondi	Diffusione crescente
Neuromorphic computing	Architettura hardware	46.9× prestazioni, 72.7× efficienza energetica	Ricerca avanzata

Tabella 5.1: Confronto tra le principali tecniche per la riduzione dell'impatto energetico nei sistemi di calcolo.

Tra le tecniche citate, a livello microarchitetturale, il Dynamic Voltage and Frequency Scaling (DVFS) consente di adattare dinamicamente la tensione di alimentazione e la frequenza di clock alle esigenze computazionali delle applicazioni: gli studi citati nel capitolo evidenziano che la combinazione di DVFS con meccanismi avanzati di gestione dei thread può portare a riduzioni del consumo energetico fino al 56%, mantenendo al contempo buoni livelli prestazionali. Invece, tecniche quali clock gating e power gating intervengono sulla riduzione della potenza dei circuiti, disabilitando le unità inattive e spegnendo selettivamente parti dell'architettura.
Il contributo della pipeline balancing è un ulteriore miglioramento che modifica dinamicamente i vincoli di ritardo degli stadi della pipeline per uniformare il consumo energetico tra le diverse unità, approccio che può condurre a riduzioni del consumo energetico fino al 46%. A livello architetturale, i sistemi multi-core eterogenei consentono di allocare i carichi computazionali sui core più appropriati dal punto di vista energetico, sfruttando core a basso consumo per compiti leggeri ed attivando unità più performanti solo se necessario.
Inoltre, l'efficienza energetica può essere migliorata con l'utilizzo di algoritmi più

efficienti e tecniche di compressione dei modelli di intelligenza artificiale, come la knowledge distillation e l'utilizzo di Small Language Models, che riducono la complessità computazionale ed il numero di parametri attivi durante l'inferenza.
Infine, approcci emergenti come il neuromorphic computing propongono una revisione più radicale dell'architettura dei sistemi di calcolo, introducendo modelli di elaborazione ispirati ai sistemi biologici.

5.7 Tecnologie verdi e politiche industriali

Il settore energetico globale è attualmente protagonista di trasformazioni profonde. Infatti, il sistema di produzione dell'energia sta evolvendo da un modello fortemente dipendente dai combustibili fossili verso uno sempre più basato sulle fonti rinnovabili. Le principali economie mondiali, dall'Asia all'Europa fino al Nord America, stanno intensificando gli investimenti nelle tecnologie a basse emissioni, perseguendo obiettivi quali la transizione verso un'industria ad emissioni nulle ed un buon posizionamento nella competizione per una nuova economia globale.
Questa tendenza è riflessa anche nei dati riguardanti gli investimenti: nel 2022, gli investimenti mondiali in nuovi impianti di generazione rinnovabile hanno rappresentato circa l'80% di quelli complessivi nel settore elettrico per un valore pari a circa 1.000 miliardi di dollari. Invece, la quota destinata alla generazione elettrica da combustibili fossili è scesa al 10%. Anche gli investimenti complessivi nelle filiere di petrolio, gas e carbone risultano in diminuzione, passando da circa 1.000 miliardi di dollari nel 2015 a circa 800 miliardi nel 2022.
Nel contesto globale, la Cina riveste un ruolo dominante come principale fornitore di tecnologie energetiche "pulite" ed esportatore della maggior parte di esse. Il Paese detiene almeno il 60% della capacità produttiva mondiale per numerose tecnologie di massa, tra cui il fotovoltaico, l'eolico e le batterie, e circa il 40% della produzione globale di elettrolizzatori. [43]

L'Europa, al contrario, si configura prevalentemente come importatore di tecnologie per l'energia pulita, sebbene presenti alcune importanti eccezioni. In particolare, mentre una quota significativa delle batterie e delle componenti dei moduli fotovoltaici è importata, l'industria europea spicca nel settore delle turbine eoliche, nel quale i produttori europei sono in grado di soddisfare integralmente la domanda interna. Analogamente, nella filiera delle pompe di calore, l'Europa conserva una posizione competitiva, pur in presenza della forte capacità produttiva cinese.

Nel dettaglio del fotovoltaico, nel 2022 i Paesi dell'Unione Europea hanno installato 41,4 GW di nuova capacità solare. Tuttavia, la produzione interna è risultata fortemente insufficiente: nello stesso anno, i produttori europei hanno fabbricato soltanto 1,7 GW di wafer, 1,37 GW di celle e 9,22 GW di moduli [44]. Di conseguenza, la produzione europea ha coperto rispettivamente solo il 4%, il 3% e il 22% del fabbisogno complessivo di impianti fotovoltaici.

Una situazione ben diversa si osserva nel settore eolico. Nel 2022, i Paesi dell'UE hanno installato 19,2 GW di nuova capacità, di cui 16,7 GW onshore e 2,5 GW offshore [45]. Per l'eolico onshore, già nel 2021 i produttori europei

avevano fabbricato 17 GW di pale e oltre 11 GW di generatori e torri, coprendo rispettivamente il 102% e il 71% del fabbisogno dell'anno successivo. Nel comparto offshore, la produzione europea ha superato ampiamente la domanda interna, con capacità produttive equivalenti al 116% per le pale, al 268% per i generatori e al 280% per torri e strutture portanti [46].

Per quanto riguarda i sistemi di accumulo energetico, nel 2021 oltre il 90% delle nuove capacità di batterie installate nell'Unione Europea era destinato al settore dei veicoli elettrici [47]. Nello stesso anno, le vendite di veicoli elettrici in Europa hanno raggiunto 2,3 milioni di unità, corrispondenti a una capacità di batterie di circa 156 GWh. A fronte di tale domanda, la capacità produttiva europea si aggirava intorno ai 60 GWh, pari a circa il 38% del fabbisogno di mercato.

Infine, nel settore delle pompe di calore, la produzione europea è prevalentemente orientata al mercato interno. Nel 2021, la capacità produttiva mondiale (esclusi i condizionatori d'aria) era pari a 120 GW, di cui circa 19 GW prodotti nell'UE. Questa quota ha consentito di coprire circa il 68% delle 2,18 milioni di nuove installazioni europee. Sebbene la Cina rimanga il principale fornitore di compressori per le pompe aria-aria, l'Europa mantiene un ruolo di primo piano nella produzione di pompe aria-acqua e terra-acqua. [48]

Nel processo di transizione energetica assume un ruolo centrale l'azione delle Istituzioni nella definizione di una strategia nazionale di lungo periodo (almeno dieci anni) che sia concreta, stabile e coerente. Una pianificazione di questo tipo è fondamentale per poter creare le condizioni necessarie allo sviluppo delle filiere industriali legate alle energie rinnovabili. Questa esigenza è stata più volte sottolineata dalle imprese coinvolte poichè la costruzione di una filiera nazionale delle rinnovabili richiede una regia centrale in grado di coordinare e governare un processo complesso. Inoltre, poiché la transizione energetica implica l'attuazione simultanea di molteplici misure, che spaziano dalla regolazione dei mercati allo sviluppo delle infrastrutture comuni, dalla conversione degli usi finali dell'energia nei settori dei trasporti, dell'industria e dei consumi domestici, fino alla definizione di strumenti di investimento pubblico. Tutti questi interventi devono essere inseriti all'interno di un quadro normativo stabile e coerente che sia capace di garantire continuità e prevedibilità nel tempo.
Al contrario, se le politiche industriali sono orientate al breve termine, introducono elementi di incertezza che tendono a scoraggiare gli investimenti strutturali.
In tali condizioni, gli operatori sono incentivati ad adottare strategie opportunistiche orientate sull'approvvigionamento da catene del valore estere, spesso più preparate a soddisfare la domanda.

In questo contesto, le imprese hanno evidenziato la necessità di una comunicazione più diretta e continuativa tra le istituzioni e i principali stakeholder della transizione energetica, in particolare per progetti altamente innovativi o caratterizzati da un'elevata intensità di capitale. Si tratta di richieste che risultano coerenti con l'impostazione delineata a livello europeo nell'ambito del Green Deal, dove si va a creare uno scenario che prevede la possibilità di ricorrere ad aiuti di Stato per lo

sviluppo di tecnologie strategiche per la transizione energetica.
All'interno di un quadro generale caratterizzato da maggiore stabilità delle politiche industriali e da un coinvolgimento più attivo delle istituzioni, le imprese hanno inoltre individuato una serie di misure specifiche a supporto dello sviluppo delle filiere verdi nazionali. Tra queste, emerge l'esigenza di accompagnare il sostegno pubblico all'investimento con interventi a favore delle comunità interessate dai nuovi insediamenti industriali che garantiscano la disponibilità di alloggi, scuole e servizi per l'infanzia, lo sviluppo urbanistico e, soprattutto, la formazione di tecnici specializzati.
Un ulteriore ambito di intervento riguarda il ruolo delle istituzioni nella comunicazione verso l'opinione pubblica dei valori e dei benefici della transizione energetica: le imprese ritengono fondamentale promuovere iniziative di informazione e sensibilizzazione che coinvolgano mondo associativo ed educativo, al fine di favorire la costruzione di consenso sociale e di una visione condivisa.
Infine, in merito al ruolo delle grandi imprese all'interno delle Comunità Energetiche Rinnovabili (CER), diversi operatori hanno espresso perplessità rispetto all'attuale impostazione europea, che considera le CER esclusivamente come strumenti a finalità sociale e ne limita quindi la partecipazione ai grandi gruppi industriali. Dunque, un coinvolgimento delle grandi imprese potrebbe costituire un importante fattore per il raggiungimento degli obiettivi di decarbonizzazione e di na crescita più rapida e strutturata delle Comunità Energetiche, grazie al contributo in termini di competenze tecniche, capacità organizzative e risorse finanziarie. [49]

Capitolo 6

Quantum Computing

6.1 Principi di funzionamento

Il quantum computing rappresenta un settore emergente dell'informatica e dell'ingegneria dei sistemi di calcolo che sfrutta i principi fondamentali della meccanica quantistica per affrontare problemi computazionali altrimenti non trattabili efficientemente dai computer classici.
Il campo del quantum computing è interdisciplinare e comprende sia lo sviluppo dell'hardware quantistico sia la progettazione di algoritmi quantistici. Sebbene la tecnologia sia ancora in fase di ricerca e sviluppo, i progressi recenti indicano che i computer quantistici su larga scala saranno in grado di risolvere problemi di elevata complessità computazionale che risultano impraticabili per i supercomputer classici, o che richiederebbero tempi di calcolo improbabili.
Grazie allo sfruttamento delle leggi della meccanica quantistica, un computer quantistico teoricamente potrebbe eseguire determinate operazioni con un'accelerazione esponenziale rispetto alle architetture classiche (vedi Figura 6.1). In particolare, problemi che richiederebbero migliaia di anni di elaborazione su un computer tradizionale potrebbero essere risolti in tempi dell'ordine di minuti o ore.
La crescita esponenziale significa raddoppiare una quantità ad ogni iterazione che porta rapidamente a valori enormi. Un esempio tipico è quello di un foglio di carta spesso un decimo di millimetro che, piegato ripetutamente su sé stesso raddoppiandone ogni volta lo spessore, dopo circa 42 piegature raggiungerebbe una distanza paragonabile a quella tra la Terra e la Luna. Un principio analogo governa l'aumento della potenza di calcolo nei sistemi quantistici.
La meccanica quantistica, che descrive il comportamento della materia e dell'energia su scale atomiche e subatomiche, introduce fenomeni controintuitivi ma fondamentali: a queste dimensioni, le leggi della meccanica classica e dell'elettromagnetismo cessano di essere sufficienti.
I computer quantistici sfruttano tali fenomeni, osservabili ad esempio in particelle come elettroni e fotoni, per implementare modelli di calcolo, memorizzazione dell'informazione e strategie matematiche di risoluzione dei problemi che non trovano un equivalente nell'informatica classica.

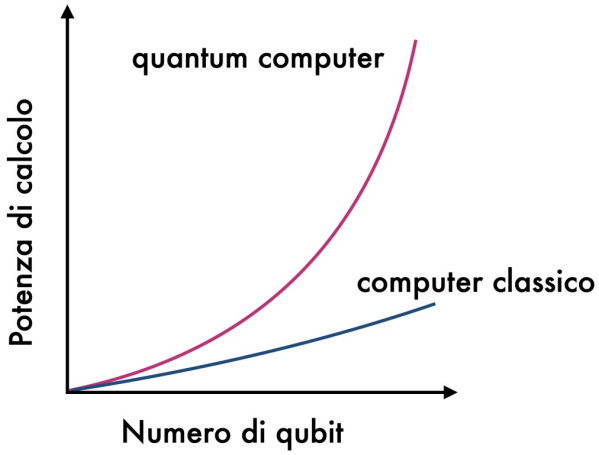

Figura 6.1: Grafico di comparazione tra computer classico e quantum computer (crescita esponenziale). [50]

L'unità fondamentale dell'informatica classica è il bit, che può assumere esclusivamente uno dei due stati possibili: 0 oppure 1, come mostrato in Figura 6.2. Questa limitazione è alla base del linguaggio binario che governa tutti i computer tradizionali, nei quali ogni informazione e ogni operazione devono essere ricondotte a sequenze di valori discreti.
I computer quantistici utilizzano invece il quantum bit o qubit, un'unità di informazione che non ha il vincolo della dicotomia binaria. Il qubit sfrutta le proprietà quantistiche della materia, in particolare, due fenomeni quantistici risultano centrali nel funzionamento di un computer quantistico: la sovrapposizione e l'entanglement. Grazie al principio di sovrapposizione, un sistema quantistico può trovarsi contemporaneamente in più stati possibili. Analogamente, un qubit non è limitato allo stato 0 o allo stato 1, ma può esistere in una combinazione continua di entrambi. Un esempio è quello di una moneta lanciata in aria: fino al momento in cui viene osservata, non è né "testa" né "croce", ma una sovrapposizione di entrambi gli stati. [50]
Dal punto di vista computazionale, la sovrapposizione consente ad un computer quantistico di eseguire molte operazioni in parallelo, aumentando drasticamente la capacità di esplorare simultaneamente numerose soluzioni possibili. Questa caratteristica rappresenta uno dei principali fattori che sta alla base della superiorità teorica dei computer quantistici rispetto a quelli classici per determinati problemi.
L'altro principio fondamentale è l'entanglement, un fenomeno per cui due o più particelle quantistiche diventano correlate in modo tale che lo stato di ciascuna non può essere descritto indipendentemente da quello delle altre, anche quando sono separate da grandi distanze. La modifica dello stato di una particella implica istantaneamente una variazione nello stato dell'altra, indipendentemente dalla distanza che le separa.
Applicato ai qubit, l'entanglement permette di costruire sistemi in cui le unità di informazione operano in forte correlazione reciproca. Questo consente dunque l'implementazione di algoritmi quantistici estremamente efficienti, in grado di rappresen-

tare e manipolare correlazioni complesse che in un computer classico richiederebbero quantità proibitive di memoria e potenza di calcolo. [51]

Figura 6.2: Illustrazione della diffirenza tra bit (computer classico) e qubit (computer quantistico). [52]

6.2 Maggiore efficienza energetica

Nel 2022, OpenAI ha rilasciato ChatGPT, un modello linguistico di grandi dimensioni in grado di generare testi coerenti e di simulare una conversazione umana con un elevato grado di accuratezza: è solo uno dei numerosi traguardi nell'ambito dell'intelligenza artificiale, un settore caratterizzato da rapidi e continui progressi. Tuttavia, il raggiungimento degli attuali livelli di accuratezza e prestazioni richiede infrastrutture computazionali ad alta intensità energetica, con conseguenti emissioni significative di gas serra. A titolo esemplificativo, il supercomputer più veloce al mondo assorbe circa 8 megawatt di potenza anche in stato di inattività, una quantità di energia sufficiente ad alimentare migliaia di abitazioni. Poiché circa l'82% del consumo energetico mondiale deriva ancora da combustibili fossili, questo aumento si traduce direttamente in un incremento delle emissioni globali di CO_2: è fondamentale ridurre le emissioni globali del 50% entro il 2030 per limitare l'aumento della temperatura media globale a 1,5 °C.
Se le attuali ambizioni computazionali non saranno accompagnate da un miglioramento dell'efficienza energetica, rischieranno di entrare in conflitto con gli obiettivi di sostenibilità climatica. Di conseguenza, si stanno esplorando nuove strategie per contenere il consumo energetico senza rinunciare però al progresso computazionale ed una delle soluzioni più discusse è rappresentata proprio dal calcolo quantistico.

Alcuni casi di studio mostrano come si stia già trovando applicazione pratica. Google, attraverso la propria divisione Quantum AI, ha esplorato l'uso di algoritmi avanzati per ottimizzare il funzionamento di parchi eolici, migliorando la previsione dei pattern di vento ed ottenendo incrementi significativi nella produzione energetica. Allo stesso modo, IBM ha avviato iniziative dedicate allo sviluppo di algoritmi quantistici per l'ottimizzazione delle reti elettriche, con l'obiettivo di ridurre le perdite e facilitare l'integrazione su larga scala delle fonti rinnovabili.

Dal punto di vista computazionale, i vantaggi teorici del calcolo quantistico in questo ambito derivano principalmente dalla capacità di risolvere problemi di ottimizzazione complessi in tempi inferiori rispetto agli approcci classici, dall'elevata accuratezza delle simulazioni fisiche e dalla scalabilità rispetto alla crescente complessità dei sistemi energetici moderni. L'elemento fondamentale per ridurre sprechi e adattarsi rapidamente a variazioni della domanda o dell'offerta energetica risiede nella possibilità di effettuare ottimizzazioni quasi in tempo reale.
Guardando al futuro, la combinazione tra calcolo quantistico ed IA appare come una delle migliori direzioni per cercare di affrontare la transizione energetica consentendo una gestione più efficiente, sostenibile ed intelligente delle risorse. Il calcolo quantistico non si configura dunque come una tecnologia isolata, ma come amplificatore delle capacità per gli strumenti computazionali già esistenti, destinato a diventare progressivamente un elemento chiave nei processi decisionali legati all'energia del futuro. [53]

Il potenziale del calcolo quantistico nel migliorare l'efficienza energetica si estende a numerosi settori industriali, con possibilità di ottimizzazione difficilmente affrontabili con il solo calcolo classico.
Nel settore delle energie rinnovabili, ad esempio, il calcolo quantistico può essere utilizzato per ottimizzare il posizionamento e la gestione di impianti eolici e fotovoltaici, tenendo conto di un numero elevato di variabili ambientali, meteorologiche ed infrastrutturali. Una previsione più accurata delle condizioni del vento o dell'irraggiamento solare consente di massimizzare la produzione di energia e ridurre gli sprechi dovuti ad una gestione non ottimale delle risorse.
Analogamente, nelle reti elettriche intelligenti, smart grid, gli algoritmi quantistici possono supportare l'ottimizzazione in tempo reale della distribuzione dell'energia, migliorando l'affidabilità del sistema, riducendo le perdite di trasmissione e facilitando l'integrazione di fonti rinnovabili caratterizzate da una produzione intermittente.
Anche il settore dei trasporti rappresenta un ambito di applicazione particolarmente promettente. Il calcolo quantistico può contribuire all'ottimizzazione delle batterie per veicoli elettrici, alla pianificazione dei percorsi logistici a minimo consumo energetico ed alla gestione del traffico urbano, con benefici diretti sia in termini di riduzione delle emissioni sia dei costi operativi.
Nel contesto della manifattura industriale, le simulazioni quantistiche consentono di progettare materiali e processi produttivi più efficienti dal punto di vista energetico, riducendo il fabbisogno di energia e le perdite associate alle fasi di trasformazione.
Inoltre, nella gestione degli edifici, sensori quantistici ed algoritmi avanzati possono essere impiegati per ottimizzare sistemi di riscaldamento, raffreddamento e illuminazione, adattandoli dinamicamente alle condizioni ambientali e all'occupazione degli spazi, con una riduzione significativa dei consumi energetici complessivi.

In specifici contesti di machine learning, ad esempio, i computer quantistici sono in grado di costruire spazi vettoriali in modo più efficiente consentendo di ridurre il numero di operazioni necessarie durante la fase di addestramento ed aumentando l'efficienza complessiva.
Nonostante ciò, il calcolo quantistico presenta ancora sfide energetiche rilevanti.

In particolare, il consumo di energia tende ad aumentare rapidamente con il numero di operazioni sequenziali, a causa della difficoltà di preservare la coerenza quantistica nel tempo. Inoltre, non esiste ancora una metrica standardizzata per valutare l'efficienza energetica dei computer quantistici, analoga ai gigaflops per watt utilizzati nel calcolo classico.
Un ulteriore aspetto critico è rappresentato dai requisiti ingegneristici: molti computer quantistici devono operare a temperature prossime allo zero assoluto, per ridurre il rumore ed il tasso di errore, con un conseguente costo energetico legato ai sistemi criogenici.
Nonostante esistano diverse tecnologie di qubit, come quelle basate su ioni intrappolati, atomi neutri o fotoni, ciascuna presenta compromessi differenti in termini di scalabilità ed efficienza energetica. È quindi improbabile che il calcolo quantistico sostituisca completamente quello classico nel breve o medio termine: rimane ancora oggetto di ricerca nella riduzione complessiva del consumo energetico; si prevede però un'integrazione ibrida, in cui i chip quantistici verranno utilizzati all'interno dei supercomputer. [54]

6.3 Limiti attuali del Quantum Computing

6.3.1 Criogenia

Il calcolo quantistico criogenico si riferisce al funzionamento dei processori quantistici a temperature estremamente basse, tipicamente inferiori a 15 millikelvin (mK). La necessità di operare in ambiente criogenico deriva dalla natura dei qubit poiché si tratta di sistemi estremamente sensibili al rumore ambientale che può causare la perdita delle proprietà quantistiche, come sovrapposizione ed entanglement, indispensabili per il calcolo. Il raffreddamento criogenico consente quindi di sopprimere il rumore termico, ridurre le interazioni indesiderate con l'ambiente e prolungare i tempi di coerenza, rendendo possibile l'esecuzione affidabile delle operazioni quantistiche.
In particolare, le basse temperature sono essenziali per abilitare la superconduttività, fenomeno grazie al quale alcuni materiali conducono corrente senza resistenza elettrica. Questa proprietà è alla base del funzionamento dei qubit superconduttori (i più diffusi) e consente di migliorare significativamente la fedeltà delle porte logiche quantistiche, riducendo i tassi di errore.
Analogamente, anche i qubit di spin, come quelli implementati in semiconduttori, beneficiano di temperature prossime allo zero assoluto, che permettono un controllo più preciso degli stati quantistici.

Diverse piattaforme di calcolo quantistico dipendono quindi dalla criogenia. I qubit superconduttori, utilizzati da aziende come IBM, Google e Rigetti, richiedono tipicamente temperature comprese tra 10 e 15 mK. I qubit di spin nei semiconduttori, sviluppati ad esempio da Intel, operano spesso a temperature inferiori a 100 mK.
Anche i qubit topologici, sebbene ancora prevalentemente a livello teorico, sono previsti funzionare in condizioni criogeniche simili per garantire la stabilità degli

stati quantistici. Al contrario, altre piattaforme, come le trappole ioniche o i qubit fotonici, non richiedono un raffreddamento così estremo, ma presentano compromessi rilevanti in termini di scalabilità, integrazione e velocità operativa.
Tali condizioni operative sono ottenute mediante sistemi di raffreddamento altamente specializzati, in particolare i refrigeratori a diluizione, e rappresentano un requisito fondamentale per l'implementazione delle principali architetture di computer quantistici.
Il centro dell'infrastruttura criogenica è composto dal refrigeratore a diluizione, un sistema a circuito chiuso che sfrutta una miscela di isotopi di elio-3 ed elio-4 per raggiungere temperature prossime allo zero assoluto. All'interno di questi apparati, il processore quantistico viene collocato nello stadio più freddo, intorno ai 10 mK, mentre l'elettronica di controllo e lettura a microonde è distribuita su stadi progressivamente più caldi, tipicamente a 1 K e 4 K. Per garantire l'isolamento dall'ambiente esterno, vengono anche implementate schermature magnetiche e sistemi di filtraggio avanzati, necessari a prevenire interferenze elettromagnetiche che potrebbero compromettere l'integrità del calcolo.
Nonostante la criogenia renda possibile l'hardware quantistico attuale, introduce anche importanti sfide ingegneristiche, la scalabilità rappresenta uno degli ostacoli principali: integrare milioni di qubit all'interno di un singolo refrigeratore a diluizione comporta limiti in termini di spazio, cablaggio e gestione termica. Inoltre, la distribuzione della potenza e dei segnali di controllo a temperature criogeniche richiede lo sviluppo di nuova elettronica capace di operare in condizioni estreme senza introdurre poi ulteriore dissipazione di calore.

Dal punto di vista energetico, i sistemi criogenici comportano consumi elevati e una manutenzione complessa, sollevando interrogativi sulla sostenibilità a lungo termine del calcolo quantistico su larga scala.
Per affrontare queste criticità, una delle direzioni di ricerca più promettenti è lo sviluppo di tecnologie Cryo-CMOS, ovvero circuiti elettronici progettati per funzionare direttamente a basse temperature. L'integrazione di tali componenti potrebbe aiutare a ridurre la complessità del cablaggio tra ambiente caldo e ambiente criogenico, migliorando l'efficienza complessiva del sistema. Sono in corso anche studi su architetture quantistiche operanti a temperatura ambiente; tuttavia si tratta di approcci che si trovano ancora in una fase preliminare e non offrono ancora prestazioni comparabili a quelle dei sistemi criogenici.
È infine importante sottolineare che la criogenia nel calcolo quantistico non è limitata al solo raffreddamento dei qubit, ma è essenziale anche per il funzionamento degli amplificatori quantistici utilizzati nella lettura degli stati, per l'instradamento e il filtraggio criogenico dei segnali a microonde e per l'hardware di interfaccia tra sistemi quantistici e classici. [55]

6.3.2 Error correction

La correzione degli errori quantistici (Quantum Error Correction, QEC) è un elemento fondamentale per la realizzazione di computer quantistici affidabili e scalabili. Si basa sull'utilizzo di codici di correzione degli errori quantistici (Quantum Error Correction Codes, QECC) progettati per proteggere gli stati quantistici dagli errori che possono insorgere durante le operazioni di calcolo o di trasmissione dell'infor-

mazione.
A differenza del calcolo classico, in campo quantistico non è possibile copiare direttamente uno stato arbitrario a causa del teorema di non clonazione; di conseguenza, i QECC distribuiscono l'informazione quantistica su più qubit fisici introducendo ridondanza. Il funzionamento di tali codici si articola tipicamente in tre fasi: la codifica dello stato quantistico, la rilevazione degli errori e la loro successiva correzione.
In questo contesto, assumono particolare rilevanza i codici di superficie che costituiscono una delle architetture più promettenti per la QEC grazie alla loro tolleranza agli errori e alla compatibilità con implementazioni fisiche bidimensionali. L'analisi di tali codici richiede una comprensione approfondita della loro struttura, delle modalità di gestione degli errori e del ruolo svolto dai parametri che ne determinano le prestazioni, tra cui i tassi di errore e il concetto di soglia di errore.
Alla base dei QEC vi sono gli operatori di Pauli X, Y e Z, le cui proprietà di commutazione e anticommutazione consentono l'identificazione degli errori. In particolare, gli operatori X e Z anticommutano, come esprime la relazione

$$(X \otimes I)(I \otimes Z) = -(I \otimes Z)(X \otimes I) \tag{6.1}$$

implicando che la misura di determinati operatori stabilizzatori permette di rivelare la presenza di errori senza disturbare l'informazione logica. In pratica, gli stabilizzatori di tipo X sono utilizzati per individuare errori di tipo Z, mentre quelli di tipo Z consentono di rilevare errori di tipo X.

Un ruolo cruciale nel processo di rilevazione degli errori è svolto dai qubit ausiliari. Questi qubit vengono accoppiati ai qubit di dati e successivamente misurati, permettendo di ottenere informazioni sugli errori presenti nel sistema senza però effettuare una misura diretta sui qubit che codificano l'informazione logica ed evitando dunque il collasso degli stati quantistici utili al calcolo. I risultati delle misure sugli ausiliari forniscono indicazioni sulla natura e sulla posizione degli errori avvenuti.
La procedura di rilevazione e correzione degli errori viene ripetuta ciclicamente attraverso i cosiddetti round di correzione, ossia un ciclo completo di misure degli stabilizzatori e di applicazione delle eventuali operazioni correttive. Il numero di round eseguiti e la loro frequenza devono essere bilanciati con la distanza del codice, poiché un numero insufficiente di round può portare alla propagazione degli errori, mentre un numero eccessivo aumenta la probabilità di introdurre nuovi errori durante le operazioni di misura.
I QECC inoltre distinguono tra qubit fisici e qubit logici: i primi corrispondono alle entità fisiche reali implementate nell'hardware quantistico, mentre gli altri rappresentano l'informazione quantistica codificata su un insieme di qubit fisici.
Un codice di correzione degli errori quantistici è spesso indicato con la notazione

$$[[n, k, \delta]],$$

dove n è il numero di qubit fisici utilizzati per la codifica, k è il numero di qubit logici protetti dal codice e δ è la distanza del codice, ovvero il numero minimo di errori fisici necessari per indurre un errore logico non correggibile.

Un parametro centrale per la valutazione delle prestazioni di un QECC è il tasso di errore fisico, che misura la probabilità di errore associata alle operazioni quantistiche come l'applicazione delle porte logiche, le misure e l'inizializzazione dei qubit: valori elevati di questo tasso indicano un'elevata frequenza di operazioni difettose, compromettendo l'affidabilità complessiva.
Invece, il tasso di errore logico quantifica la probabilità che un errore si manifesti a livello dei qubit logici, nonostante l'impiego di codici di correzione. Grazie all'azione dei QECC, il tasso di errore logico dovrebbe essere molto inferiore a quello fisico. Tuttavia, se il tasso di errore fisico risulta troppo elevato o se la distanza del codice è insufficiente, alcuni errori possono comunque propagarsi e tradursi in errori logici. In questo contesto assume un ruolo centrale il concetto di soglia di errore (threshold) che rappresenta il valore massimo del tasso di errore fisico al di sotto del quale un codice di correzione è in grado di ridurre efficacemente il tasso di errore logico aumentando la distanza del codice. Quando il tasso di errore fisico supera tale soglia, la ridondanza introdotta dal codice non è più sufficiente a compensare la quantità di errori nel sistema, rendendo inefficace la correzione. Per questo motivo, gran parte degli studi sulla QEC si concentra su regimi operativi al di sotto della soglia, fornendo indicazioni per la progettazione, l'ottimizzazione e la scalabilità dei futuri computer quantistici. [56]

6.4 Confronto qualitativo dei consumi: classico vs quantistico

Al centro della crescente attenzione all'impatto energetico delle tecnologie di calcolo avanzate, in un momento storico in cui l'efficienza energetica è un parametro critico tanto quanto la potenza computazionale, i sistemi quantistici rappresentano un caso interessante poiché richiedono infrastrutture complesse e ad alto consumo energetico, in particolare per il raffreddamento criogenico necessario al mantenimento della coerenza quantistica. Il computer quantistico Sycamore di Google ha attirato l'attenzione della comunità scientifica quando nel 2019 è riuscito a risolvere un problema computazionale molto complesso in circa 200 secondi anziché 10.000 anni come avrebbe richiesto all'incirca il supercomputer classico più potente disponibile all'epoca. Questa differenza di prestazioni mostra un vantaggio in specifiche classi di problemi, come l'ottimizzazione combinatoria, la crittografia, la simulazione di materiali e di sistemi quantistici. Settori industriali quali la logistica, la scoperta di nuovi farmaci e la modellazione finanziaria possono trarre benefici significativi da queste capacità, in particolare grazie all'elevata riduzione dei tempi di calcolo.

Dal punto di vista hardware, un processore classico può contenere miliardi di transistor, mentre i computer quantistici più avanzati raggiungono al massimo qualche migliaio di qubit. Nonostante questo confronto possa far apparire i sistemi quantistici ancora limitati, è fondamentale sottolineare, come illustrato precedentemente, che i qubit scalano in modo esponenziale. Di conseguenza, anche poche centinaia di qubit ben controllati sono sufficienti per affrontare problemi che risultano intrattabili per i sistemi classici.
Dal punto di vista architetturale, può apparire controintuitivo il fatto che le porte

logiche classiche operino su scale temporali dell'ordine dei picosecondi, mentre le operazioni quantistiche avvengono tipicamente su scale di nanosecondi. Tuttavia, la velocità di una singola operazione non è l'unico parametro rilevante. Il calcolo quantistico introduce un paradigma diverso: mentre un computer classico elabora le informazioni in modo sequenziale entro limiti ben definiti, un computer quantistico è in grado di esplorare un numero esponenziale di possibilità in un singolo passo computazionale, cambiamento di paradigma che risulta particolarmente rilevante per applicazioni di ottimizzazione, simulazione e riconoscimento di pattern complessi.
Una delle principali limitazioni attuali del calcolo quantistico è rappresentata dagli alti tassi di errore poichè, a differenza dei transistor classici, estremamente stabili, i qubit sono soggetti al fenomeno della decoerenza, che comporta una rapida perdita dell'informazione quantistica.

È sorta l'ipotesi di un computer quantistico da 256 qubit che rappresenta un salto concettuale dal momento che la quantità di informazione classica che un sistema di questo tipo potrebbe elaborare supera il numero di atomi nell'universo osservabile.
Un ambito emblematico di applicazione è la chimica quantistica, uno dei problemi computazionali più complessi esistenti: la simulazione di interazioni molecolari, la previsione di reazioni chimiche e lo studio del comportamento atomico richiedono risorse computazionali enormi. I computer classici affrontano questi problemi tramite approcci approssimati o metodi di forza bruta, il cui costo cresce esponenzialmente con il numero di atomi coinvolti. Poiché i sistemi chimici obbediscono intrinsecamente alle leggi della meccanica quantistica, il calcolo quantistico risulta naturalmente più adatto a descriverli, rendendo possibili simulazioni altrimenti impraticabili. [57]

Nel confronto energetico tra calcolo classico e quantistico, è comune adottare un'impostazione sperimentale in cui problemi algoritmici di complessità limitata vengono eseguiti su differenti architetture di calcolo. L'esecuzione degli stessi algoritmi su più sistemi quantistici, ed in condizioni temporali diverse, permette di evidenziare non solo le differenze prestazionali tra dispositivi, ma anche la loro variabilità nel tempo, influenzata da ricalibrazioni, aggiornamenti hardware e degradazione dei qubit, caratteristiche tipiche delle piattaforme quantistiche attuali.
I problemi considerati in questo tipo di confronto sono generalmente semplici e ben definiti così da ridurre l'impatto di fattori esterni e isolare le proprietà intrinseche dell'hardware quantistico, scelta che consente di analizzare l'influenza del quantum volume, e quindi della profondità e della complessità del circuito, su grandezze fondamentali quali il tempo di esecuzione, il consumo energetico e l'affidabilità del risultato.
Per quanto riguarda il consumo energetico dei computer quantistici invece, la misura diretta sull'hardware fisico risulta generalmente infattibile. Di conseguenza, l'energia consumata viene stimata a partire dal tempo di esecuzione, assumendo una potenza media costante del sistema coerente con le specifiche dei produttori in modo che tempo di esecuzione ed energia risultino direttamente proporzionali,

rendendo il tempo una variabile chiave per l'analisi energetica e permettendo di concentrare l'attenzione sulla misura temporale e sulla valutazione statistica degli output.
A causa del rumore, della decoerenza e delle imperfezioni nelle operazioni di gate, ogni circuito viene eseguito molte volte ed i risultati vengono interpretati in termini statistici. La success rate, definita come la frazione di output corretti rispetto al numero totale di esecuzioni, rappresenta quindi un indicatore essenziale della qualità del calcolo.
In generale, il tempo di esecuzione mostra variazioni significative tra sistemi quantistici diversi anche a parità di problema. Di grande rilevanza è anche l'aumento della complessità circuitale che comporta di conseguenza un incremento del consumo energetico.
Eseguendo problemi diversi sul medesimo dispositivo, le differenze di consumo energetico risultano invece generalmente contenute suggerendo che l'hardware ha un ruolo più determinante rispetto alla natura del problema. Si evidenzia dunque una correlazione diretta tra quantum volume e consumo energetico: sistemi con quantum volume più elevato tendono a supportare circuiti più profondi, aumentando sia il tempo di esecuzione sia l'energia richiesta. Inoltre, uno stesso dispositivo può fornire risultati migliori o peggiori in momenti diversi, in funzione di ricalibrazioni, degrado dei qubit o progressiva dismissione dell'hardware confermando che non esiste un sistema quantistico che risulti sistematicamente superiore in tutti i contesti: allo stato attuale della tecnologia, non emergono indicazioni di un vantaggio energetico del calcolo quantistico rispetto a quello classico per problemi di bassa complessità.

Nel caso del calcolo classico, il consumo energetico può essere misurato direttamente sull'hardware. A parità di problemi, i tempi di esecuzione risultano estremamente ridotti e poco sensibili alla natura dell'algoritmo, mentre il consumo energetico presenta variazioni contenute ma misurabili. A differenza dei sistemi quantistici attuali, non si osserva una proporzionalità diretta tra tempo di esecuzione ed energia consumata, poiché il consumo nei computer classici dipende principalmente dal carico computazionale effettivo e non dall'infrastruttura di supporto.
Il confronto tra i due paradigmi mostra che, per problemi di bassa complessità, il calcolo classico risulta di gran lunga più veloce ed energeticamente efficiente. I sistemi quantistici richiedono tempi di esecuzione molto più elevati e consumano ordini di grandezza in più di energia. Potrà risultare giustificato solo per problemi di complessità estremamente elevata, per i quali il calcolo classico diventa impraticabile: il vero valore del calcolo quantistico risiede nella sua integrazione strategica con il calcolo classico. [58]

6.5 Applicazioni del Quantum Computing per ridurre l'impatto energetico

6.5.1 Simulazioni chimiche

I computer quantistici possono essere impiegati per generare dati estremamente accurati sul comportamento degli elettroni in atomi, molecole e materiali, dati

che sarebbero estremamente costosi da ottenere con metodi classici. Questi dati "quantum-accurate" possono essere utilizzati per addestrare modelli di AI eseguiti su computer classici, capaci di predire rapidamente le proprietà di sistemi simili. In questo modo si combina l'accuratezza del calcolo quantistico con la velocità e la scalabilità dell'AI, rendendo accessibili simulazioni che oggi sono limitate da costi computazionali insostenibili.
Questo concetto viene efficacemente descritto attraverso la metafora della Jacob's Ladder, introdotta dal fisico teorico John P. Perdew, per rappresentare la gerarchia dei metodi computazionali utilizzati nella chimica quantistica e nella scienza dei materiali. Alla base della scala si trovano modelli puramente classici, in cui gli atomi sono trattati come particelle connesse da molle: si tratta di modelli molto rapidi, in grado di simulare milioni di atomi su tempi lunghi, ma con un livello di precisione limitato. Salendo, si incontrano metodi semiempirici, che introducono una prima descrizione quantistica, e successivamente approcci come il metodo di Hartree–Fock e la Density Functional Theory, nei quali il comportamento quantistico degli elettroni è descritto esplicitamente, mentre le interazioni elettroniche vengono trattate in modo mediato. Questi metodi rappresentano uno standard nella simulazione di materiali, ma il loro costo computazionale ne limita l'applicazione a sistemi di dimensioni ridotte, tipicamente dell'ordine di poche centinaia di atomi.
Ai gradini più alti della Jacob's Ladder si collocano i metodi più accurati, come il coupled-cluster e la full configuration interaction (vedi Figura 6.3), che forniscono una descrizione estremamente fedele della correlazione elettronica. Tuttavia, tali metodi incontrano rapidamente un "muro esponenziale" di complessità computazionale: il numero di configurazioni elettroniche cresce in maniera esplosiva con il numero di elettroni rendendo questi approcci impraticabili su computer classici per sistemi che vadano oltre molecole molto piccole. È proprio in questo regime che il calcolo quantistico promette di intervenire: poiché i computer quantistici operano secondo le stesse leggi della meccanica quantistica che governano i sistemi elettronici, essi sono adatti a simulare sistemi fortemente correlati, nei quali gli elettroni non possono essere trattati come entità indipendenti.

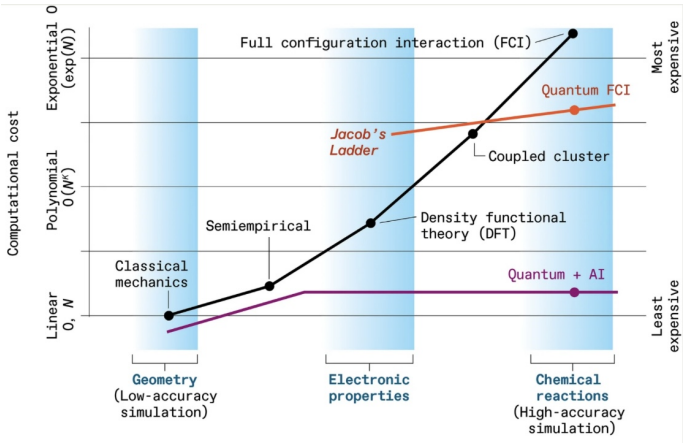

Figura 6.3: Grafico comparativo del costo computazionale in relazione all'accuratezza delle simulazioni. [59]

Tuttavia, poiché l'hardware quantistico rimane limitato e costoso, il passo successivo è sfruttare questi risultati per addestrare modelli di intelligenza artificiale. L'AI agisce quindi come un emulatore: una volta addestrata su dati di qualità quantistica, è in grado di riprodurre predizioni di livello quasi quantistico ad un costo computazionale nettamente inferiore. In questo senso, l'AI "appiattisce" la curva costo-accuratezza della Jacob's Ladder, consentendo di effettuare predizioni rapide e affidabili su computer classici.
Questa sinergia tra calcolo quantistico ed AI apre la strada a nuove prospettive per la chimica e la scienza dei materiali: dalla progettazione di nuovi catalizzatori alla scoperta di materiali per batterie più sicure ed efficienti, dalla rimozione di inquinanti allo sviluppo di nuovi farmaci, la possibilità di combinare accuratezza quantistica e velocità computazionale promette di trasformare profondamente il modo in cui vengono affrontati problemi complessi. Quindi, il vero passo avanti non deriva dal calcolo quantistico isolato, ma dalla sua integrazione strategica con l'intelligenza artificiale, capace di rendere fruibili e scalabili informazioni che fino ad oggi sono rimaste confinate ai livelli più alti e inaccessibili della Jacob's Ladder. [59]

6.5.2 Ottimizzazione quantistica

La maggior parte dei sistemi di intelligenza artificiale più avanzati, dalle auto a guida autonoma ai modelli di previsione dei mercati finanziari, opera costantemente adattandosi, perfezionandosi e apprendendo dall'esperienza. Alla base di queste decisioni si trova il processo di ottimizzazione: ogni modello di apprendimento automatico, che si tratti di riconoscere immagini, classificare segnali o prevedere un valore futuro, deve individuare la soluzione migliore all'interno di uno spazio di possibilità tendenzialmente enorme. Con l'aumento della complessità dei modelli e delle dimensioni dei dataset, questa ricerca diventa però più onerosa dal punto di vista computazionale.
In questo contesto si inserisce l'ottimizzazione quantistica, un approccio emergente che sfrutta i principi dell'informatica quantistica probabilistica per affrontare problemi di ottimizzazione complessi.

L'ottimizzazione quantistica è il risultato di decenni di ricerca interdisciplinare tra fisica e informatica. Le sue origini risalgono alla fine degli anni Novanta con lo studio della meccanica quantistica. Nel tempo, queste idee si sono evolute in una famiglia di algoritmi progettati per sfruttare gli effetti quantistici nella ricerca efficiente di soluzioni in spazi di grande dimensionalità.
Alla base dell'ottimizzazione quantistica vi sono tre elementi fondamentali: i qubit, gli algoritmi quantistici e i circuiti quantistici. Gli algoritmi forniscono la struttura logica che guida l'esplorazione dello spazio delle soluzioni, mentre i circuiti quantistici, composti da sequenze di porte quantistiche, orchestrano l'evoluzione del sistema, modulando le interazioni tra i qubit per indirizzare il calcolo verso una soluzione a bassa energia, interpretata come una soluzione ottimale o quasi ottimale del problema.
Il confronto tra approcci classici e quantistici mette in luce differenze profonde. I metodi classici tendono a valutare una soluzione alla volta, con risultati deterministici e riproducibili, e funzionano bene per problemi di dimensioni limitate o moderatamente complesse. Gli approcci quantistici, invece, operano in modo

probabilistico e sono progettati per affrontare problemi di scala molto maggiore, nei quali il numero di configurazioni cresce esponenzialmente ed il risultato di una singola esecuzione non è necessariamente definitivo, ma viene raffinato attraverso ripetizioni e analisi statistiche.
Dal punto di vista operativo, un problema di ottimizzazione ha come obiettivo quello di individuare il minimo globale energetico. Il problema viene quindi codificato in un circuito quantistico e, attraverso un processo controllato di evoluzione del sistema, il computer quantistico guida lo stato verso regioni di energia sempre più bassa. Una volta raggiunto uno stato stabile, la misurazione dei qubit fornisce una soluzione approssimata, che viene spesso ulteriormente raffinata con l'intervento di algoritmi classici, dando luogo ad un approccio ibrido.

Una delle tecniche più studiate in questo ambito è la ricottura quantistica (o quantum annealing), ispirata ad un processo fisico in cui un materiale viene raffreddato lentamente per raggiungere uno stato di minima energia. Analogamente, la ricottura quantistica accompagna il sistema verso la soluzione ottimale, ossia la ricerca di un minimo globale, risultando particolarmente adatta a problemi di ottimizzazione combinatoria come la pianificazione, l'instradamento o l'allocazione delle risorse. I risultati, essendo probabilistici, richiedono più esecuzioni per migliorarne l'affidabilità.
Un altro approccio rilevante è rappresentato dal Quantum Approximate Optimization Algorithm, che affronta problemi combinatori alternando due operatori energetici: uno che codifica l'obiettivo ed i vincoli del problema e un altro che favorisce l'esplorazione dello spazio delle soluzioni. Si tratta di un metodo che si inserisce in un contesto ibrido in cui il computer quantistico genera ipotetiche soluzioni ed un computer classico ottimizza iterativamente i parametri dell'algoritmo.
In modo complementare, il Variational Quantum Eigensolver (VQE) si concentra su problemi di ottimizzazione continua ed è particolarmente utilizzato in chimica e nella scienza dei materiali per stimare lo stato fondamentale di sistemi quantistici. Grazie alla struttura variazionale e ai requisiti contenuti in termini di qubit, il VQE risulta adatto agli attuali dispositivi quantistici rumorosi.
Inoltre, sono in fase di studio anche algoritmi quantistici per la programmazione semidefinita, con applicazioni nell'apprendimento automatico, nell'elaborazione dei segnali e nei sistemi di controllo. Stanno ulteriormente emergendo algoritmi di ottimizzazione ispirati ai quanti, eseguibili su hardware classico, che cercano di replicare alcune strategie quantistiche in attesa di computer quantistici più maturi.

Nonostante i rapidi progressi, l'adozione su larga scala dell'ottimizzazione quantistica è ancora limitata da diversi fattori. Gli attuali sistemi quantistici dispongono di un numero ridotto di qubit, sono influenzati da rumore e decoerenza e la simulazione classica di grandi sistemi quantistici è estremamente costosa, rendendo difficile testare e validare algoritmi complessi.
Ciò nonostante, l'ottimizzazione quantistica, pur non essendo ancora una soluzione matura per applicazioni su larga scala, offre una visione concreta di come la combinazione di calcolo quantistico e classico possa, in futuro, affrontare problemi di ottimizzazione attualmente considerati intrattabili. [60]

6.6 Sfide del Quantum Computing e prospettive future

La sfida più rilevante del calcolo quantistico è con ogni probabilità la decoerenza dei qubit. Infatti, essendo estremamente sensibili all'ambiente circostante, anche perturbazioni minime con l'ambiente esterno come fluttuazioni termiche, rumore elettromagnetico o vibrazioni meccaniche, possono causare la perdita delle loro proprietà quantistiche. Questo fenomeno, noto appunto come decoerenza, compromette la capacità del sistema di mantenere stati quantistici coerenti per tempi sufficientemente lunghi da consentire computazioni affidabili. Il controllo della decoerenza richiede di conseguenza lo sviluppo di nuovi materiali, l'adozione di tecniche di fabbricazione avanzate ed un'esplorazione approfondita di differenti approcci architetturali al calcolo quantistico.
In aggiunta, lo sviluppo dell'hardware quantistico presenta difficoltà rilevanti: la realizzazione di qubit di alta qualità e dell'elettronica di controllo associata è un processo complesso, che coinvolge tecnologie avanzate ed infrastrutture altamente specializzate. Esistono numerose piattaforme di qubit, ciascuna caratterizzata da specifici vantaggi e limitazioni e l'individuazione di una tecnologia che sia al tempo stesso scalabile, stabile e tollerante ai guasti è proprio uno degli obiettivi principali della ricerca contemporanea. [61]

Le difficoltà del calcolo quantistico non sono limitate esclusivamente all'hardware. Anche dal punto di vista algoritmico, i computer quantistici impongono un cambio di paradigma rispetto al calcolo classico poichè sono intrinsecamente più complessi e richiedono di affrontare i problemi computazionali in modi radicalmente diversi, sfruttando fenomeni quali la sovrapposizione e l'entanglement.
Dunque, come spiegato precedentemente, la correzione degli errori quantistici emerge come una delle sfide centrali: i computer quantistici sono altamente vulnerabili al rumore e agli errori causati dalle interazioni con l'ambiente che tendono ad accumularsi nel tempo e a degradare progressivamente la qualità delle computazioni.
Inoltre, nonostante i dispositivi quantistici attuali abbiano dimostrato prestazioni promettenti in specifici compiti, il numero di qubit disponibili rimane limitato rispetto alle esigenze di un computer quantistico universale. Incrementare il numero di qubit fino a centinaia o migliaia, mantenendo però elevati tempi di coerenza e bassi tassi di errore, costituisce un'altra delle principali sfide tecnologiche del settore.
Invece, il software quantistico si trova ancora in una fase embrionale e l'assenza di un ecosistema software maturo costituisce un ulteriore limitazione per l'adozione del calcolo quantistico anche su larga scala. Vi è una forte necessità di sviluppare nuovi linguaggi di programmazione, compilatori e strumenti di ottimizzazione capaci di tradurre problemi reali in circuiti quantistici efficienti, tenendo conto anche delle limitazioni fisiche dei dispositivi attuali.

Un aspetto strettamente correlato riguarda le interfacce tra sistemi classici e quantistici. I computer quantistici non sono destinati a sostituire quelli classici,

bensì ad integrarli come acceleratori per determinate classi di problemi. Risulta quindi fondamentale sviluppare metodi affidabili ed efficienti per lo scambio di dati e il coordinamento delle operazioni tra i due paradigmi di calcolo.
Con la progressiva maturazione del settore, emerge inoltre l'esigenza di standard e protocolli condivisi per l'hardware, il software e le interfacce di comunicazione: è essenziale per garantire compatibilità tra piattaforme diverse. A ciò si aggiunge il problema del benchmarking, poiché la valutazione oggettiva e comparabile delle prestazioni dei computer quantistici è ancora in una fase iniziale e priva di metriche universalmente accettate.
Infine, tutte le sfide descritte contribuiscono a rendere il calcolo quantistico estremamente costoso: il personale altamente qualificato, l'hardware specializzato e le catene di approvvigionamento complesse comportano investimenti significativi, ponendo interrogativi sulla sostenibilità economica nel breve e medio termine. [**?**]

6.6.1 Quantum machine learning

Una delle possibili prospettive future più promettenti è il Quantum Machine Learning (QML) che utilizza le capacità di elaborazione dell'informazione integrando le tecniche del machine learning con gli algoritmi quantistici. Il QML sfrutta le proprietà della meccanica quantistica per affrontare problemi computazionali complessi caratterizzati da correlazioni non lineari difficilmente individuabili mediante tecniche di Machine Learning e Deep Learning tradizionali. L'idea alla base del QML consiste nell'utilizzare la capacità dei computer quantistici di memorizzare ed elaborare simultaneamente un'enorme quantità di informazioni, con l'obiettivo di migliorare le prestazioni di apprendimento, ottimizzazione e danalisi dei dati rispetto agli approcci classici.
Questa superiorità deriva dalla possibilità di analizzare dataset di grandi dimensioni in tempi molto inferiori rispetto ai metodi convenzionali classici. Grazie ai qubit che possono trovarsi in una sovrapposizione di stati, i computer quantistici sono in grado di eseguire molte operazioni in parallelo, consentendo un'elevata accelerazione dei calcoli utile per algoritmi di apprendimento automatico che richiedono l'elaborazione simultanea di grandi quantità di dati o la valutazione di più soluzioni.
Un secondo elemento è la capacità di ottimizzazione avanzata: gli algoritmi quantistici possono affrontare problemi di ottimizzazione combinatoria, spesso intrattabili per i metodi classici, in modo più efficiente come, per esempio, l'ottimizzazione di portafogli finanziari, la pianificazione delle risorse, la progettazione di materiali e la risoluzione di problemi complessi ad alta dimensionalità.
Il QML si distingue inoltre per la sua efficacia nella gestione di dati complessi e ad alta dimensionalità. La rappresentazione quantistica dei dati consente di codificare informazioni complesse in spazi di Hilbert ad alta dimensione offrendo nuove modalità di analisi per dataset di grandi dimensioni, approccio che apre prospettive innovative in ambiti come la bioinformatica, la medicina computazionale e la ricerca scientifica basata su grandi volumi di dati.
Si parla di vantaggio quantistico quando un computer è in grado di eseguire calcoli in un tempo nettamente inferiore rispetto a quello richiesto da un supercomputer classico, arrivando in alcuni casi a ridurre tempi computazionali da miliardi di anni a pochi minuti o secondi, ma anche e soprattutto per la vantaggiosa capacità di

affrontare problemi che risultano impraticabili per gli algoritmi classici. [62]

Le applicazioni del Quantum Machine Learning sono tante e trasversali. In ambito finanziario, gli algoritmi quantistici possono analizzare i mercati in modo estremamente rapido, individuando pattern nascosti e con una gestione del rischio in tempo reale. Nel settore della medicina e della ricerca farmaceutica, il QML può accelerare la scoperta di nuovi farmaci, migliorare la modellazione di molecole complesse e supportare diagnosi più accurate attraverso l'analisi di grandi quantità di dati genetici e clinici. Nell'industria, trova applicazione nell'ottimizzazione dei processi produttivi e nella progettazione di sistemi complessi, contribuendo ad incrementare l'efficienza operativa. In ambito di sicurezza informatica, l'impiego di algoritmi quantistici favorisce lo sviluppo di sistemi di difesa capaci di individuare e contrastare minacce in modo più rapido ed efficace.
Infine, l'applicazione del calcolo quantistico all'apprendimento apre prospettive promettenti nel miglioramento delle strategie decisionali e dei processi di ottimizzazione.
Il Quantum Machine Learning può inoltre contribuire allo sviluppo di tecniche di computer vision più efficienti, migliorando le prestazioni degli algoritmi di Deep Learning per l'elaborazione, il riconoscimento e la segmentazione delle immagini.

L'integrazione tra intelligenza artificiale e calcolo quantistico promette di ampliare significativamente i confini delle capacità computazionali attuali grazie al parallelismo quantistico. Tuttavia, il QML deve ancora confrontarsi con sfide rilevanti: l'hardware quantistico è tuttora in fase sperimentale e presenta limitazioni in termini di stabilità, scalabilità ed accessibilità. La progettazione di algoritmi di Quantum Machine Learning è complessa e richiede competenze altamente specializzate, che combinano informatica, fisica quantistica e matematica avanzata. Inoltre, la decoerenza e gli errori quantistici rappresentano un ostacolo significativo poiché la fragilità dei qubit può compromettere l'affidabilità delle computazioni e limitarne l'applicabilità pratica degli algoritmi.
Nonostante queste difficoltà, il Quantum Machine Learning rappresenta una delle frontiere più promettenti, gli investimenti in ricerca e sviluppo sono in costante crescita ed i progressi scientifici suggeriscono che, nel prossimo decennio, potremmo assistere ad un'adozione sempre più diffusa di soluzioni ibride classico-quantistiche in cui computer classici e quantistici opereranno in parallelo, ciascuno specializzato in specifiche tipologie di compiti. [63]

Conclusioni

In questo lavoro è stato indagato il rapporto tra lo sviluppo dell'intelligenza artificiale ed il crescente fabbisogno energetico delle infrastrutture digitali che ne supportano il funzionamento. L'evoluzione degli algoritmi di apprendimento automatico e la sempre maggiore complessità dei modelli, in particolare nel campo del deep learning, hanno determinato negli ultimi anni un aumento significativo delle risorse computazionali richieste sia nella fase di addestramento sia nella fase di utilizzo dei sistemi di IA.
Inoltre, l'analisi delle architetture hardware e delle infrastrutture di calcolo ha evidenziato come il consumo energetico dei sistemi di intelligenza artificiale sia strettamente correlato alla struttura fisica delle piattaforme computazionali su cui tali algoritmi vengono eseguiti.
Dispositivi come GPU, TPU ed altre unità di elaborazione specializzate sono progettati per accelerare le operazioni matematiche delle reti neurali profonde, incremento di prestazioni che però viene spesso accompagnato da un aumento significativo della potenza assorbita. Di conseguenza, la crescente diffusione delle applicazioni basate su IA comporta un'espansione dei data center ed un aumento della domanda globale di energia elettrica.
Infatti, dal punto di vista ambientale, questo fenomeno solleva importanti questioni legate alla sostenibilità delle tecnologie digitali: l'incessante funzionamento dei data center, l'elevato fabbisogno energetico dei sistemi di raffreddamento e la produzione di hardware specializzato contribuiscono massivamente alle emissioni globali di gas serra e al consumo di risorse naturali. L'impatto ambientale dell'intelligenza artificiale non si limita quindi alla sola fase di utilizzo degli algoritmi, ma coinvolge l'intero ciclo di vita delle tecnologie digitali, dall'addestramento all'inferenza, dalla produzione dei semiconduttori fino allo smaltimento dei dispositivi elettronici.
Allo stesso tempo, è emerso come l'intelligenza artificiale possa rappresentare anche uno strumento estremamente utile per migliorare l'efficienza energetica e la sostenibilità di applicazioni basate su algoritmi di machine learning per ottimizzare la gestione delle reti elettriche, migliorare le previsioni della produzione da fonti rinnovabili, ridurre i consumi energetici ed ottimizzare i processi industriali.
Nella ricerca scientifica, inoltre, le applicazioni nel campo della fisica delle particelle, dell'astrofisica e delle onde gravitazionali dimostrano come gli algoritmi di apprendimento automatico possano accelerare i processi di analisi dei dati e favorire l'individuazione di segnali e correlazioni difficilmente identificabili con metodi tradizionali. Dunque, l'IA sta assumendo un ruolo sempre più centrale nell'analisi di grandi quantità di dati sperimentali e nella simulazione di sistemi fisici complessi non limitandosi a svolgere un ruolo di supporto computazionale, ma diventando strumento capace di contribuire alla scoperta di nuova scienza.

Alla luce di queste considerazioni, emerge chiaramente come il futuro sviluppo dell'intelligenza artificiale richieda un equilibrio tra innovazione tecnologica ed efficienza energetica. La progettazione di algoritmi più efficienti, l'adozione di architetture hardware ottimizzate, lo sviluppo di modelli di dimensioni più contenute e l'utilizzo crescente di tecniche di on-device rappresentano alcune delle strategie più promettenti per ridurre l'impatto energetico.
Anche soluzoni emergenti come il neuromorphic computing ed il Quantum Computing potrebbero aprire la strada a modelli computazionali diversi da quelli attuali in grado di offrire nuove opportunità per una maggiore efficienza energetica dei sistemi di elaborazione dell'informazione.
Tuttavia, queste tecnologie si trovano ancora in una fase iniziale di sviluppo per cui presentano numerose sfide tecniche ed ingegneristiche che dovranno essere affrontate nei prossimi anni.

In conclusione, l'intelligenza artificiale rappresenta una delle tecnologie più comode, influenti e promettenti del nostro tempo, ma il suo sviluppo futuro dovrà confrontarsi necessariamente con le sfide legate alla sostenibilità energetica ed ambientale. Solo attraverso un approccio integrato che unisca innovazione tecnologica, efficienza computazionale e responsabilità ambientale, sarà possibile garantire che il progresso dell'IA contribuisca positivamente allo sviluppo scientifico, economico e sociale della società contemporanea.

Appendice

Non è possibile stimare con precisione il consumo reale di ChatGPT nella stesura di questa tesi poichè dipende dall'infrastruttura dei data center e dai modelli usati (es. GPT-3 o GPT-4), dati che non sono pubblici. Tuttavia, è possibile fare una stima indiretta basata su medie di settore e letteratura (assumendo che una singola query a un LLM consumi circa 2–3 Wh per richiesta):

$$E_{\text{tot}} = N \cdot E_{\text{query}}, \tag{6.2}$$

dove N è il numero di richieste effettuate all'AI e $E_{\text{query}} \approx 2.5$ Wh.

Ad esempio, ipotizzando $N = 100$ richieste:

$$E_{\text{tot}} = 100 \cdot 2.5 \text{ Wh} = 250 \text{ Wh} = 0.25 \text{ kWh}. \tag{6.3}$$

Si tratta di un'approssimazione dell'energia totale associata alle inferenze che include: calcolo GPU nel data center, raffreddamento e infrastruttura server.
In termini di consumo domestico, una lampadina da 60 W accesa per 1 ora consuma 60 Wh [64].
Quindi, 0,25 kWh corrispondono a:

$$\text{Ore lampadina} = \frac{0,25 \text{ kWh}}{0,06 \text{ kWh/h}} \approx 4,17 \text{ ore}$$

Bibliografia

[1] International Data Corporation (IDC). Idc - market intelligence and technology research. https://www.idc.com, 2026.

[2] Web Crew. Una grande quantità di dati. https://webcrew.it/una-grande-quantita-di-dati/, 2020.

[3] Renato Schirripa. Processori nell'era dell'i.a.: differenza tra cpu, gpu, tpu, npu, lpu, 2025. Link all'articolo.

[4] Redazione ZeroUno. Oltre la cpu: differenze, vantaggi e limiti di gpu, npu, dpu e tpu nell'ai. https://www.zerounoweb.it/techtarget/searchdatacenter/oltre-la-cpu-differenze-vantaggi-e-limiti-di-gpu-npu-dpu-e-tpu-nellai/, 2026.

[5] Digitech Bytes. Arcades of ai: Gpu vs npu workloads on laptops. https://digitechbytes.com/emerging-consumer-tech-explained/gpu-vs-npu-workloads/, 2025.

[6] The Purple Struct. Cpu vs gpu vs tpu vs npu: Ai hardware architecture guide 2025. https://www.thepurplestruct.com/blog/cpu-vs-gpu-vs-tpu-vs-npu-ai-hardware-architecture-guide-2025, 2025.

[7] Massimo Nannini. Il costo energetico degli algoritmi di intelligenza artificiale, 2024. Automation Technology Magazine: link all'articolo.

[8] Saverio Lapini. Quanti kg di co2 emette un kwh elettrico in italia? https://ollum.it/blog/quanti-kg-co2-per-kwh-elettrico-italia/, 2025.

[9] Jens Gröger, Felix Behrens, Peter Gailhofer, and Inga Hilbert. Environmental impacts of artificial intelligence: Evaluation of current trends and compilation of an overview study, 2025. Link all'articolo.

[10] Greenpeace Italia. Il volto nascosto dell'ia: vorace di risorse e fonte di inquinamento, 2025. Link all'articolo.

[11] International Energy Agency. Energy and ai: Energy demand from ai, 2025. Link all'articolo.

[12] David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. Carbon

emissions and large neural network training. *CoRR*, abs/2104.10350, 2021. Link all'articolo.

[13] Alessandro Trizio. Quanta energia divora l'ai? https://eywadivulgazione.it/ai-consumo-energetico/, 2024.

[14] Federica Arnaud. Perché il deep learning funziona? https://deeplearningitalia.com/perche-il-deep-learning-funziona/, 2024.

[15] Istituto Nazionale di Fisica Nucleare (INFN). Rivoluzione ia: l'impatto dell'intelligenza artificiale sulla ricerca e la società. https://www.infn.it/rivoluzione-ia/, 2025.

[16] Md Ashraful Haque. Revolutionizing physics: The role of artificial intelligence in modern scientific discoveries. *British Journal of Physics Studies*, 3(1):12–21, 2025. Accessed: 10 March 2026.

[17] Le Scienze. L'intelligenza artificiale "alza il volume" dell'universo: "ora potremo ascoltarlo meglio". https://www.lescienze.it/comunicati-stampa/2025/09/05/news/intelligenza_artificiale_volume_universo_onde_gravitazionali-19975451/, September 2025. Comunicato stampa.

[18] Sirui Wu, Nicola R. Napolitano, Crescenzo Tortora, Rodrigo von Marttens, Luciano Casarini, Rui Li, and Weipeng Lin. Total and dark mass from observations of galaxy centers with machine learning. *Astronomy & Astrophysics (A&A)*, 686:A80, 2024. Link all'articolo.

[19] WhiteFiber. Optimizing ai models for efficiency. https://www.whitefiber.com/blog/optimizing-ai-models, 2025. Blog article.

[20] Glennis Sullivan. Efficient accuracy trade-off. https://medium.com/@yolov8architecture/efficient-accuracy-trade-off-6a86fa6a9d35, Aug 2025. Medium article.

[21] Jeremy Enos, Craig Steffen, Joshi Fullop, Michael Showerman, Guochun Shi, Kenneth Esler, Volodymyr Kindratenko, John E. Stone, and James C. Phillips. Quantifying the impact of gpus on performance and energy efficiency in hpc clusters. In *Proceedings of the International Conference on Green Computing*, pages 317–324, 2010.

[22] Orbyta. Recommender systems: principali metodologie degli algoritmi di suggerimento. https://orbyta.it/insights/recommender-systems-algoritmi-di-suggerimento/, 2025. Articolo informativo sulle principali tecniche dei sistemi di raccomandazione.

[23] Ultralytics. Sistema di raccomandazione. https://www.ultralytics.com/it/glossary/recommendation-system, 2025. Glossario online.

[24] Epsilon. Pressroom. https://www.epsilon.com/us/about-us/pressroom, 2026. Pagina delle notizie e comunicati stampa di Epsilon.

[25] Natalie Severt. An introduction to recommender systems (+9 easy examples). https://www.iteratorshq.com/blog/an-introduction-recommender-systems-9-easy-examples/, September 2025. Blog article.

[26] URMET S.p.A. Assistenti vocali: cosa sono e come funzionano nelle smart home. https://www.urmet.com/it-it/Professionista/Notizie-ed-Eventi/assistenti-vocali-smart-home, February 2025. Articolo informativo sul ruolo degli assistenti vocali nella smart home.

[27] Mostafa Ibrahim. Ai in our day-to-day health and fitness. https://www.ultralytics.com/it/blog/ai-in-our-day-to-day-health-and-fitness, August 2024. Blog post.

[28] IMA Financial Group. Ai and autonomous ships: Redefining risk in marine insurance. https://imacorp.com/insights/insurance-insights-ai-and-autonomous-ships-redefining-risk-in-marine-insurance, May 2025. Articolo informativo sull'impatto dell'intelligenza artificiale e delle navi autonome nell'assicurazione marittima.

[29] Coursera Staff. On-device ai: Powering the future of computing. https://www.coursera.org/articles/on-device-ai, May 2025. Articolo informativo su AI on-device e vantaggi rispetto al cloud.

[30] Elsevier. European community. https://www.sciencedirect.com/topics/social-sciences/european-community, 2026. ScienceDirect Topics.

[31] Irene Niet, Laura Van den Berghe, and Rinie van Est. Societal impacts of ai integration in the eu electricity market: The dutch case. *Technological Forecasting and Social Change*, 192:122554, 2023.

[32] PCMag Editors. Why being polite to chatgpt is costing openai millions and wasting electricity. https://tech.yahoo.com/articles/why-being-polite-chatgpt-costing-150624055.html, April 2025. Articolo informativo su costi energetici e operativi legati all'uso di ChatGPT.

[33] Sustainability Directory. Digital rebound effect: un fenomeno di sostenibilità digitale. https://climate.sustainability-directory.com/term/digital-rebound-effect/, 2026. Glossario online sul fenomeno del Digital Rebound Effect.

[34] Abhishek Sinkar, Hadi Ghasemi, Michael Schulte, Ulya R. Karpuzcu, and Nam Sung Kim. Low-cost per-core voltage domain support for power-constrained high-performance processors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(4):747–758, 2014. Link all'articolo.

[35] Charles Leech and Tom J. Kazmierski. Energy efficient multi-core processing. *Electronics*, 18(1):3–10, 2014. Link all'articolo.

[36] Statisticseasily. Cos'è l'efficienza dell'algoritmo: una panoramica dettagliata. https://it.statisticseasily.com/glossario/cos%27-l%

27efficienza-dell%27algoritmo%3A-una-panoramica-dettagliata/, 2025. Glossario su efficienza degli algoritmi con riferimenti alla complessità temporale e spaziale.

[37] Sustainability Directory. Efficient algorithms. https://energy.sustainability-directory.com/term/efficient-algorithms/, 2025. Definizione e spiegazione di algoritmi efficienti con riferimento all'uso delle risorse in contesti di sostenibilità energetica.

[38] Dave Bergmann. Cos'è la distillazione della conoscenza (knowledge distillation). https://www.ibm.com/it-it/think/topics/knowledge-distillation, 2026. Articolo pubblicato su IBM Think da Dave Bergmann con panoramica sulla tecnica di knowledge distillation nel machine learning.

[39] Rina Diane Caballar. Che cosa sono i modelli linguistici di piccole dimensioni (small language models). https://www.ibm.com/it-it/think/topics/small-language-models, 2026. Articolo pubblicato su IBM Think da Rina Diane Caballar con panoramica sui Small Language Models (SLM).

[40] Kyle Dickman. Neuromorphic computing: the future of ai. https://www.lanl.gov/media/publications/1663/1269-neuromorphic-computing, March 2025. Articolo su 1663 Magazine pubblicato dal Los Alamos National Laboratory.

[41] Peter Hess. What is neuromorphic or brain-inspired computing? https://research.ibm.com/blog/what-is-neuromorphic-or-brain-inspired-computing, October 2024. Blog post su IBM Research.

[42] IBM. What is edge ai? https://www.ibm.com/think/topics/edge-ai, 2026. Articolo pubblicato su Think — IBM.

[43] International Energy Agency. Energy technology perspectives 2023. https://www.iea.org/reports/energy-technology-perspectives-2023, 2023. Rapporto pubblicato dall'International Energy Agency, OECD Publishing, Paris. Cfr. p. 58.

[44] SolarPower Europe. Eu market outlook for solar power 2023-2027. https://www.solarpowereurope.org/insights/outlooks/eu-market-outlook-for-solar-power-2023-2027, 2023. Rapporto pubblicato da SolarPower Europe con analisi del mercato solare europeo del 2023 e prospettive per il 2027.

[45] WindEurope. Wind energy in europe: 2023 statistics and the outlook for 2024-2030. https://windeurope.org/data/product/wind-energy-in-europe-2023-statistics-and-the-outlook-for-2024-2030/, 2024. Rapporto pubblicato da WindEurope (Febbraio 2024). Dati su installazioni e capacità eolica in Europa nel 2023.

[46] International Energy Agency. World energy outlook 2023. https://www.iea.org/reports/world-energy-outlook-2023, 2023. Rapporto pubblicato dall'International Energy Agency (IEA), OECD Publishing, Paris.

[47] Marek Bielewski, Andreas Pfrang, Silvia Bobba, Aleksandra Kronberga, and et al. Clean energy technology observatory: Batteries for energy storage in the european union – 2022 status report on technology development, trends, value chains and markets. https://publications.jrc.ec.europa.eu/repository/handle/JRC130724, 2022. Rapporto tecnico del Joint Research Centre, Commissione Europea.

[48] Giovanni Sgaravatti, Simone Tagliapietra, and Cecilia Trasi. Cleantech manufacturing: where does europe really stand? https://www.bruegel.org/analysis/cleantech-manufacturing-where-does-europe-really-stand-0, May 2023. Pubblicato da Bruegel. Data di pubblicazione: 17 maggio 2023.

[49] Confindustria. Le imprese italiane e la competitività nelle tecnologie verdi. https://public.confindustria.it/repository/2025/03/27015924/Le-imprese-italiane-e-la-competitivita-nelle-tecnologie-verdi_vf.pdf, March 2025. Rapporto pubblicato da Confindustria.

[50] Natalia Milazzo and Sergio Cima. Quantum computing: strumenti di calcolo che vengono dal futuro. https://orizzonti.polito.it/it/2025/quantum-computing-strumenti-di-calcolo-che-vengono-dal-futuro/, September 2025. Articolo pubblicato su Orizzonti Polito (Politecnico di Torino).

[51] Josh Schneider and Ian Smalley. Cos'è il quantum computing? https://www.ibm.com/it-it/think/topics/quantum-computing, 2026. Articolo pubblicato su IBM Think con spiegazione generale del quantum computing, comprensivo di principi fondamentali e applicazioni.

[52] Istituto Nazionale di Fisica Nucleare. Computer quantistici. https://www.infn.it/fisica/big-data-calcolo-e-quantum-computing/computer-quantistici/, 2026. Pagina informativa su computer quantistici pubblicata dall'INFN.

[53] Sophia Chen. Are quantum computers really energy efficient? *Nature Computational Science*, 3:457–460, 2023. Link all'articolo.

[54] Meegle. Quantum computing in quantum energy efficiency. https://www.meegle.com/en_us/topics/quantum-computing-applications/quantum-computing-in-quantum-energy-efficiency/, 2026. Articolo su applicazioni del quantum computing per l'efficienza energetica, inclusi sistemi di ottimizzazione, smart grids e simulazioni avanzate.

[55] SpinQuanta. What is cryogenic quantum computing and why it matters. https://www.spinquanta.com/news-detail/what-is-cryogenic-quantum-computing-and-why-it-matters, May

2025. Articolo informativo su cryogenic quantum computing, con spiegazione delle temperature criogeniche per mantenere la coerenza dei qubit e della tecnologia necessaria.

[56] Avimita Chatterjee, Subrata Das, and Swaroop Ghosh. Q-pandora unboxed: Characterizing resilience of quantum error correction codes under biased noise. *Applied Sciences*, 15(8):4555, 2025. Link all'articolo.

[57] Bao Tran. Quantum computing vs. classical computing: Speed and performance stats. https://patentpc.com/blog/quantum-computing-vs-classical-computing-speed-and-performance-stats/, February 2026. Articolo informativo pubblicato su PatentPC con confronto tra quantum computing e classical computing in termini di velocità e prestazioni.

[58] Elena Desdentado, Coral Calero, M. Angeles Moraga, Manuel Serrano, and Felix Garcia. Exploring the trade-off between computational power and energy efficiency: An analysis of the evolution of quantum computing and its relation to classical computing. *Journal of Systems and Software*, 217:112165, 2024. Link all'articolo.

[59] Chi Chen and Matthias Troyer. How quantum data can teach ai to do better chemistry. https://spectrum.ieee.org/quantum-chemistry, March 2026. Articolo su *IEEE Spectrum* che discute del vasto potenziale del quantum computing applicato alla chimica computazionale e alla scoperta di farmaci e materiali.

[60] Abirami Vina. From bits to qubits: How quantum optimization is reshaping ai. https://www.ultralytics.com/blog/from-bits-to-qubits-how-quantum-optimization-is-reshaping-ai, October 2025. Articolo che esplora l'ottimizzazione quantistica e il suo impatto sull'intelligenza artificiale, descrivendo algoritmi quantistici e approcci ibridi.

[61] Matt Swayne. What are the remaining challenges of quantum computing? https://thequantuminsider.com/2023/03/24/quantum-computing-challenges/, April 2024. Articolo su sfide aperte nel quantum computing, inclusi decoerenza, error correction, scalabilità, hardware e software.

[62] OVHcloud. Cos'è il quantum machine learning. https://www.ovhcloud.com/it/learn/what-is-quantum-machine-learning/, 2026. Articolo informativo su applicazioni e potenzialità del Quantum Machine Learning, incluso l'uso di QML per accelerare e migliorare modelli di Machine Learning tradizionali.

[63] MTS Informatica. Quantum machine learning: L'incontro tra intelligenza artificiale e quantum computing. https://www.mtsinformatica.com/ai/quantum-machine-learning-lincontro-tra-intelligenza-artificiale-e-quantum-compu February 2025. Articolo introduttivo su Quantum Machine Learning, che descrive come l'intelligenza artificiale e il calcolo quantistico convergano per affrontare problemi complessi e nuove applicazioni.

[64] U.S. Department of Energy. Energy saver guide: Tips on saving money and energy at home. https://www.energy.gov/energysaver/energy-saver-guide-tips-saving-money-and-energy-home, 2022.