POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

Mercato delle Criptovalute e Intelligenza Artificiale

Relatore
prof. Franco Varetto

Candidato
Eugenio Valenti

Anno Accademico 2020/2021

INDICE

INTRODUZIONE **1**

1. EFFICIENZA DEL MERCATO AZIONARIO **3**

1.1 TIPOLOGIE DI EFFICIENZA 3

1.2 EFFICIENT MARKET HYPOTHESIS 5

1.3 EVIDENZE EMPIRICHE A FAVORE DELL'EMH 10

1.4 PUZZLES 19

1.5 RAZIONALITÁ LIMITATA E BOLLE SPECULATIVE 22

1.6 CONSIDERAZIONI RIEPILOGATIVE 26

2. CRIPTOVALUTE **27**

2.1 DEFINIZIONE DI CRIPTOVALUTA 29

2.2 CARATTERISTICHE DELLE CRIPTOVALUTE E BLOCKCHAIN 35

2.3 STRUTTURA DEL MERCATO DELLE CRIPTOVALUTE 42

2.4 PRINCIPALI CRIPTOVALUTE 44

2.5 CENTRAL BANK DIGITAL COIN 50

2.5.1 CBDC NEGLI STATI UNITI 52

2.5.2 CBDC IN CINA 53

2.5.3 CBDC IN EUROPA 54

3. ARTIFICIAL INTELLIGENCE, MACHINE LEARNING E DEEP LEARNING **55**

3.1 ARTIFICIAL INTELLIGENCE 55

3.2 MACHINE LEARNING 58

3.3 DEEP LEARNING E RETI NEURALI 63

3.3.1 MODELLO MATEMATICO DEL NEURONE ARTIFICIALE 66

3.3.2 SIMPLE E DEEP NEURAL NETWORK 71

3.3.3 ADDESTRAMENTO DI UNA RETE NEURALE 74

3.3.4 TIPOLOGIE DI RETI NEURALI 77

3.3.5 RETI NEURALI RICORRENTI 79

4. EFFICIENZA DEL MERCATO DELLE CRIPTOVALUTE **86**

4.1 MODELLO IN PYTHON 88

4.1.1 DATI 88

4.1.2 ARCHITETTURA DELLA RETE NEURALE 93

4.1.3 ADDESTRAMENTO 94

4.1.4 RISULTATI 97

5. CONCLUSIONE **104**

BIBLIOGRAFIA E SITOGRAFIA **107**

INTRODUZIONE

Il tema delle criptovalute si presenta come un argomento relativamente nuovo, sebbene la prima criptovaluta, per come sono considerate attualmente, sia stata creata poco più di un decennio fa. Il mercato associato agli scambi effettuati in criptovalute può essere, perciò, definito giovane e poco maturo, a differenza del mercato azionario, le cui borse valori più importanti al mondo hanno avuto origine tra il XVIII e il XIX secolo.

Data questa sostanziale differenza, risulta di particolare interesse la comprensione del grado di efficienza del mercato delle criptovalute. Se da un lato, per quanto riguarda il mercato azionario, sono stati effettuati numerosi studi in letteratura relativamente all'efficienza di tale mercato, dall'altro lato, per quanto riguarda il mercato delle criptovalute, analisi che focalizzano la loro attenzione su questo tema risultano ancora poco presenti, proprio a causa della giovinezza del marcato in questione.

L'efficienza dei mercati finanziari, di per sé, è un argomento sul quale si è dibattuto ampiamente nel corso del secolo scorso e ancora più assiduamente con il verificarsi della crisi del 2008, che ha coinvolto la maggior parte dei mercati a livello mondiale e che ha dato origine proprio all'introduzione del concetto di criptovalute, le quali hanno assunto col tempo una posizione di contrasto con il sistema finanziario tradizionale.

Nel presente lavoro di tesi, al fine di esaminare l'efficienza del marcato delle criptovalute viene adoperato un metodo innovativo, consistente nella creazione e nell'addestramento di una rete neurale ricorrente, la quale viene applicata a due indici del mercato azionario, ovvero EuroStoxx50 e S&P500, e alla coppia BTC/USD, con l'obiettivo di prevederne le quotazioni future dei prezzi. Misurando gli errori commessi dai risultati ottenuti sugli indici azionari e sul prezzo del bitcoin, è possibile ottenere un'indicazione dell'efficienza del mercato delle criptovalute in relazione a quella del mercato azionario.

Nel primo capitolo del presente lavoro di tesi, viene indagato il tema dell'efficienza del mercato azionario, tramite un'analisi degli studi che sono stati compiuti da illustri economisti, concentrando l'attenzione sull'EMH, ovvero la teoria dell'efficienza dei mercati formulata da Eugene Fama e riconosciuta come una delle teorie maggiormente accreditate nel panorama di studio in questione, e dunque analizzando approfonditamente l'efficienza informativa nelle sue diverse forme.

Il secondo capitolo ha lo scopo di esplorare, tramite una panoramica generale, il mondo delle criptovalute, partendo dalla storia delle loro origini, passando per le definizioni che sono state proposte per questo termine, analizzando dunque la tecnologia che ne permette il funzionamento e che differenzia questi nuovi asset da quelli tradizionalmente conosciuti, per poi esaminare alcuni concetti chiave, che rappresentano l'ecosistema del mercato delle criptovalute, e le prime cinque criptovalute per capitalizzazione al momento della scrittura del presente lavoro di tesi, concludendo infine con un approfondimento sul tema delle CBDC, valute digitali emesse dalle banche centrali che prendono ispirazione proprio dalle criptovalute.

Nel terzo capitolo, viene presentato il tema delle reti neurali, ovvero una branca dell'intelligenza artificiale, che permette all'essere umano di addestrare una macchina per l'apprendimento di diverse funzionalità, utili in un'ampia gamma di applicazioni. Nello specifico, dopo un breve ma doveroso excursus relativo all'intelligenza artificiale e al machine learning, e dopo una trattazione delle caratteristiche generali delle reti neurali, vengono analizzate con maggior riguardo le reti neurali ricorrenti e i relativi sviluppi.

Il quarto capitolo funge da collante per i capitoli precedenti, in quanto mira ad analizzare l'efficienza del mercato delle criptovalute in relazione a quella del mercato azionario attraverso l'implementazione di una rete neurale programmata nel linguaggio di programmazione Python, da applicare ai tre asset sopra menzionati al fine di prevedere l'evoluzione di breve periodo delle relative quotazioni.

1. EFFICIENZA DEL MERCATO AZIONARIO

L'efficienza del mercato azionario è un argomento su cui si è dibattuto molto, fin dagli inizi del secolo scorso, quando nel 1900 il giovane matematico Louis Bachelier scrisse la sua tesi di dissertazione di dottorato dal titolo "Thèorie de la Spèculation", un argomento atipico per il contesto accademico di quel periodo. Nel corso del XX secolo, la teoria del matematico francese venne presa in considerazione come punto di partenza da cui sviluppare nuove teorie relative l'efficienza dei mercati. Sono stati prodotti numerosi scritti, tra cui saggi, articoli e libri, che trattano dell'argomento in questione, sia in maniera critica, apportando prove confutative ed esempi empirici, sia a favore di questa teoria, inserendo nuovi particolari e prove argomentative a sostegno, come dimostrano gli studi compiuti da illustri economisti ed accademici quali Paul Samuelson, Alfred Cowles, Eugene Fama, Michael Jansen, Kenneth French, Harry Markowitz, Fischer Black, Myron Scholes, Robert Merton, Burton Malkiel e molti altri.
Dal punto di vista della teoria economica generale, il concetto di efficienza è stato prevalentemente identificato con l'ipotesi di concorrenza perfetta, caratterizzata da assenza di costi di transazione ed altre imperfezioni, come teorizzato nel XVIII secolo dal padre del liberismo Adam Smith ed in seguito dalla corrente macroeconomica Neoclassica.

1.1 TIPOLOGIE DI EFFICIENZA

Al fine di comprendere in modo più approfondito cosa si intenda per efficienza dei mercati, è necessario operare una suddivisione riguardante le diverse accezioni insite in tale concetto. Esistono infatti tre categorie di efficienza:

- Efficienza allocativa-funzionale: riguarda la capacità di trasferire le risorse finanziarie ed i rischi in modo da soddisfare gli obiettivi dei diversi operatori presenti sul mercato. Dato un certo livello dei prezzi, ogni operatore scambia i propri titoli al fine di assicurarsi nuove risorse e di minimizzare la propria esposizione ai vari tipi di rischio esistenti. Dunque, l'efficienza allocativa-funzionale misura quanto il mercato riesca a soddisfare i bisogni degli operatori in un'ottica di Pareto efficienza: quando non è più possibile aumentare la soddisfazione di un operatore senza peggiorare quella degli altri, allora non vi è più necessità di ulteriori scambi.
  Per come viene definita, all'interno di questa categoria rientrano altri due tipi di efficienza:
  - Fundamental-valuation efficiency: riguarda l'accuratezza della valutazione che il mercato esegue sul prezzo delle azioni. Il mercato stabilisce il prezzo delle azioni, in modo che in esso sia contenuto il rischio e il potenziale dell'attività sottostante, ovvero l'influsso dei rendimenti attesi in futuro.
  - Full-insurance efficiency: riguarda la capacità del mercato nel permettere agli operatori di assicurarsi autonomamente la fornitura di beni in ogni circostanza futura, sia tramite lo smobilizzo delle proprie attività, sia impegnandosi a consegnare determinate attività nel caso si verifichino determinati "stati di natura". I contratti per determinati beni in determinati "stati di natura" sono chiamati, nella teoria economica, contratti Arrow-Debreu.
- Efficienza tecnico-operativa: riguarda la capacità del mercato di creare strutture che permettano la presenza di un elevato numero di operatori, i quali siano in grado di effettuare ordini di acquisto o vendita in modo agevole, in tempi brevi, nel momento a loro

congeniale, al prezzo voluto, per le quantità richieste e con ridotti costi di transazione (ovvero l'insieme di costi necessari per effettuare lo scambio: costi di ricerca della controparte, costi di raccolta delle informazioni, commissioni, spread denaro-lettera, slippage, ecc…). Questo tipo di efficienza viene valutato in base ai seguenti parametri:

- Immediatezza: riguarda la possibilità per l'operatore, presente sul mercato, di effettuare uno scambio di una certa quantità, ad un certo costo, nell'esatto momento in cui desidera farlo. Lo slippage, ossia la differenza che intercorre fra prezzo al quale viene inoltrato l'ordine di acquisto o di vendita e il prezzo a cui l'ordine viene effettivamente eseguito, è un indicatore di tale parametro.
- Ampiezza: riguarda la capacità del mercato di attrarre ingenti volumi di ordini in acquisto e vendita per ogni livello di prezzo. Un mercato sufficientemente ampio dovrebbe limitare il market impact, ovvero l'effetto sui prezzi derivante da un singolo ordine.
- Spessore o profondità: riguarda la capacità del mercato di offrire una determinata quantità di un certo strumento finanziario a più prezzi vicini tra loro. Un mercato sufficientemente profondo garantirebbe, perciò, la presenza di numerosi ordini in acquisto e vendita intorno al prezzo corrente, assorbendo così gli squilibri di domanda e offerta, ovvero riducendo la presenza di vuoti di domanda e offerta e attenuando il rischio di eccessive variazioni di prezzo. Lo spread denaro-lettera o bid-ask è un indicatore dello spessore del mercato, ma può essere anche interpretato come livello di liquidità del mercato stesso, con cui si intende il tempo richiesto ad un operatore per vendere un asset sul mercato e renderlo liquido, ossia monetizzarlo.
- Elasticità: riguarda la capacità del mercato di riflettere nel prezzo variazioni, anche infinitesimali, delle quantità domandate ed offerte.

Questi quattro parametri sono intrinsecamente correlati tra loro, poiché l'incremento di uno di essi comporta il miglioramento degli altri, e viceversa.

- Efficienza informativa (information-arbitrage efficiency): riguarda la capacità del mercato di raccogliere, elaborare e diffondere le informazioni rilevanti per gli operatori che effettuano una scelta riguardo ai titoli sui quali allocare la loro ricchezza. Tale decisione avviene in condizioni di incertezza, che può essere descritta come la partizione del futuro in un insieme di stati di natura alternativi, al cui verificarsi l'operatore assegna una distribuzione di probabilità. Il ruolo delle informazioni è quello di intervenire nel processo di formulazione delle probabilità dei diversi stati di natura o nella loro revisione, e quindi nella modifica degli ordini precedentemente effettuati.

  Se si dissemina il mercato di segnali, facilmente disponibili e senza costi o con costi limitati, se ne può dedurre il contenuto informativo dall'aggiustamento dei prezzi, che incorporano le informazioni a seguito delle decisioni effettuate dagli operatori. L'intensità degli effetti, la durata del periodo lungo il quale si hanno gli aggiustamenti e il tempo che il mercato impiega per elaborare il contenuto informativo dei segnali dipendono dal grado di efficienza del mercato stesso. Dunque, dal punto di vista informativo, un mercato è efficiente se tutte le informazioni sono trasmesse perfettamente (tutti ricevono l'informazione), completamente (tutti ricevono l'intera informazione), istantaneamente (tutti ricevono l'informazione in una volta sola), e senza costi (tutti ricevono l'informazione gratuitamente).

Le tre categorie di efficienza, appena trattate, sono interconnesse tra loro, in quanto un mercato efficiente dal punto di vista tecnico-operativo permette un'adeguata formazione dei prezzi, i quali, rispecchiando le informazioni detenute dagli operatori, contribuiscono ad aumentare l'efficienza informativa, che a sua volta permette agli operatori di allocare in modo efficiente le proprie risorse, trovando le controparti per le transazioni e bilanciando la propria esposizione al rischio.

1.2 EFFICIENT MARKET HYPOTHESIS

La teoria principale e più conosciuta, riguardante l'efficienza del mercato, viene attribuita all'economista statunitense Eugene Fama, il quale pubblicò sul Journal of Finance, nel 1970, l'articolo intitolato "Efficient Capital Markets: A Review of Theory and Empirical Work". Come si può intuire dal titolo dell'opera, egli ha basato il suo elaborato sugli studi compiuti, tra il 1900 e il 1965, da altri illustri esponenti del mondo delle scienze economiche, partendo proprio dalle considerazioni effettuate ad inizio secolo da Louis Bachelier.

Il lavoro di Louis Bachelier, dal titolo "Thèorie de la Spèculation", fu il primo trattato probabilistico riferito alle fluttuazioni dei prezzi degli strumenti finanziari. In particolare, l'analisi venne realizzata prendendo come campione le obbligazioni emesse dallo Sato francese. Con l'obiettivo di sviluppare formule per determinare il prezzo dei titoli, Louis Bachelier elaborò un modello che paragonava il mercato delle obbligazioni ad un gioco equilibrato, ovvero un gioco in cui è possibile conseguire con la medesima probabilità un risultato positivo o negativo, equiparabile ad un aumento dei prezzi delle obbligazioni o ad una loro riduzione. Inoltre, egli sostenne che non era possibile prevedere le future fluttuazioni dei corsi obbligazionari, in quanto non si conoscevano con certezza tutte le informazioni relative alla politica economica dello Stato, e che dunque non si poteva avere la sicurezza che una variazione nel prezzo delle obbligazioni sarebbe avvenuta con verso positivo o negativo, in quanto i prezzi percorrono una traiettoria aleatoria definita "cammino casuale", che non dipende dagli eventi passati, ma esclusivamente dalle nuove informazioni che giungono al mercato. Sottolineando che i prezzi dispongono della proprietà di non memoria, Louis Bachelier dedusse che le variazioni dei prezzi formassero una serie di variabili casuali indipendenti ed identicamente distribuite, in quanto, se disposte in un diagramma in funzione di un periodo di riferimento temporale, assumerebbero la forma di una distribuzione Normale con media nulla e varianza unitaria. In conclusione, Louis Bachelier, nella sua tesi, sostenne che il mercato, inconsapevolmente, obbedisce ad una legge che lo governa, ovvero la legge della probabilità.

Eugene Fama si inserisce in questo filone economico con un più rigoroso approccio statistico-matematico ed una maggiore forza nell'esposizione, inaugurando una generazione di esaltazioni di tale paradigma teorico e delle sue implicazioni empiriche. Si tratta di una nuova rivoluzione per la finanza, tanto che Fama è riconosciuto come il padre della finanza moderna ed è stato insignito del premio Nobel per l'economia nel 2013 insieme a Lars Peter Hansen e Robert James Shiller.

La nozione di efficienza del mercato è strettamente legata all'ipotesi di mercato efficiente (EMH o Efficient Market Hypothesis), sviluppata da Eugene Fama, secondo cui un mercato finanziario è efficiente se i prezzi riflettono sempre pienamente le informazioni disponibili.

Acquirenti e venditori rivelano le loro informazioni in modo indiretto, attraverso le operazioni che effettuano in acquisto o vendita e tramite gli aggiustamenti di tali offerte sulla base del comportamento degli altri operatori, modificando le quantità o i prezzi a cui sono disposti ad

acquistare o vendere. Con questo processo, il mercato aggrega tutte le informazioni e convoglia questo elevato contenuto informativo all'interno dei prezzi, favorendo così la price discovery.
L'ipotesi formulata da Eugene Fama non va però confusa con il fatto che il prezzo delle azioni debba necessariamente, in ogni momento, esprimere in modo corretto il valore dell'impresa a cui si riferisce. Infatti, tra il prezzo e il valore effettivo potrebbe esserci una differenza, a causa del fatto che la natura dell'informazione, che il prezzo incorpora, non è limitata ai fondamentali economici dell'impresa, ma comprende notizie finanziarie, economiche, politiche e sociali di contorno, che possono impattare in modo rilevante sulla realtà dell'impresa stessa. Inoltre, bisogna, soprattutto, tenere in considerazione il modo in cui gli operatori presenti sul mercato percepiscono questa varietà di informazioni e come le adoperano per andare ad allocare le loro risorse sul mercato azionario. Perciò, dal momento in cui tutti questi fattori partecipano alla formazione dei prezzi, è comprensibile come tra prezzo e valore effettivo possa esserci un gap, la cui presenza, dunque, non deve compromettere il concetto di efficienza del mercato.
Chiarito il fatto che il prezzo di un'azione non corrisponda necessariamente al suo valore effettivo, può essere ora compreso il vero significato dell'affermazione di Eugene Fama: i prezzi incorporano le informazioni disponibili senza distorsioni sistematiche. Ciò significa che, anche se i prezzi debbano, secondo l'ipotesi formulata, riflettere tempestivamente e in modo corretto tutte le informazioni, non si può escludere a priori che ci possa essere la necessità di attendere un intervallo di tempo, più o meno ampio, affinché i prezzi si aggiustino attorno al loro valore effettivo.
Come conseguenza dell'EMH, in un mercato efficiente risulta, allora, impossibile stimare in modo sistematico l'andamento della quotazione dei prezzi azionari. Dunque, gli operatori non possono elaborare strategie che permettano loro di ottenere rendimenti superiori alla media complessiva di mercato, data dall'andamento dei principali indici azionari, in modo consistente nel tempo. Poiché le azioni vengono sempre scambiate ad un valore che tende ad assestarsi attorno al valore di equilibrio, gli operatori, in media, non riusciranno ad acquistare azioni quotate ad un prezzo sottovalutato per poi rivenderle quando risulteranno sopravvalutate al fine di ottenere guadagni in eccesso rispetto al tasso normale di profitto. In altre parole, un operatore, a seguito di una transazione in acquisto o vendita, potrebbe ottenere un profitto nel breve periodo, ma, a lungo termine, il rendimento della sua operazione non sarà diverso dal rendimento medio di mercato, in quanto tutte le informazioni disponibili sono già incorporate nei prezzi e poiché l'elevato numero di operatori assicurerebbe che i prezzi si assestino su valori di equilibrio.
L'ipotesi di mercato efficiente formulata da Eugene Fama, secondo cui i prezzi riflettono prontamente l'informazione disponibile, può essere espressa in un modo ancora più comprensibile, affermando che l'informazione, che può influenzare il prezzo futuro di uno strumento finanziario, viene riflessa nel prezzo attuale e non in quello futuro. Questo concetto è stato esplicato attraverso un esempio, che può risultare molto attuale, dall'economista statunitense Burton Malkiel, nel suo paper del 2011, dal titolo "The Efficient Market Hypothesis and the Financial Crisis", in cui spiega che, se un'azienda farmaceutica, le cui azioni hanno un valore di $20, ricevesse l'approvazione per la produzione di un nuovo farmaco, che aumenterebbe il valore dell'azienda a $40 nel giorno successivo, allora il prezzo dell'azione diventerebbe subito pari a $40. Inoltre, poiché qualunque acquisto del titolo ad un prezzo inferiore a $40 produrrebbe un profitto immediato, ci si può coerentemente aspettare che gli operatori facciano offerte di acquisto fino a poco prima che il prezzo raggiunga il valore atteso. Similmente a quanto analizzato in precedenza, ovviamente, è possibile che l'effetto della nuova informazione, restando nell'ambito dell'esempio apportato da Burton Malkiel, non venga immediatamente e del tutto compreso dagli operatori. Inoltre, è anche

probabile che le vendite e i profitti derivati dal nuovo farmaco non vengano stimati con precisione e, dunque, il valore apportato dalla nuova informazione può essere suscettibile di una grande varietà di stime. Poiché tra i diversi valori stimati vi è un eguale probabilità che questi siano soggetti ad una reazione eccessivamente positiva così come eccessivamente negativa, allora si conclude che non è possibile ottenere un rendimento sistematicamente positivo da questo genere di operazioni. Dunque, l'efficienza del mercato azionario porta inevitabilmente all'impossibilità per gli operatori di generare extra-profitti. Tale concetto può essere riassunto tramite un aneddoto riportato da Burton Malkiel, all'interno del suo libro dal titolo "A Random Walk Down Wall Street", scritto nel 1973, che recita così: il professore di finanza sta camminando nel campus con il suo assistente di ricerca, il quale dice: "Professore, vedo una banconota da venti dollari sul marciapiede. Vado a prenderla?", e il professore risponde: "No, certo che no, se fosse davvero lì, sarebbe già stata presa." Una versione meno estrema del racconto, come afferma lo stesso Burton Malkiel, farebbe in modo che il professore consigliasse al suo assistente di prendere immediatamente la banconota perché altrimenti lo avrebbe fatto qualcun altro. Infatti, in un mercato efficiente, la concorrenza non permette la realizzazione di rendimenti positivi in maniera regolare. Questo aneddoto può essere sintetizzato, affermando che il mercato "sconta" tutto, ovvero, come sostiene Eugene Fama, il mercato ha la capacità di incorporare tutta l'informazione disponibile nei prezzi, non lasciando opportunità di guadagno agli operatori.
Per analizzare l'affermazione, secondo cui l'informazione è pienamente incorporata nei prezzi, dal punto di vista statistico-matematico, Eugene Fama fa riferimento a tre modelli:

- Modello del Fair Game: partendo dal concetto di excess returns, definiti come

$$\varepsilon_{t+1} = \left\{\frac{(P_{t+1} - P_t)}{P_t}\right\} - \left\{\frac{(E[P_{t+1}|\Phi_t] - P_t)}{P_t}\right\} = \frac{(P_{t+1} - E[P_{t+1}|\Phi_t])}{P_t}$$
$$= r_{t+1} - E[r_{t+1}|\Phi_t]$$

  dove Φ_t indica l'insieme delle informazioni disponibili agli operatori al tempo t, P_t il prezzo al tempo t ed r_t il rendimento al tempo t, con questo approccio l'aspettativa sul rendimento atteso in eccesso viene specificata direttamente, senza alcuna considerazione sul modello che genera i rendimenti attesi, e risulta essere

$$E[\varepsilon_{t+1}|\Phi_t] = 0$$

  Ciò significa che, se Φ_t è pienamente utilizzato ed incorporato nei prezzi attesi, allora nel lungo periodo gli scostamenti tra i rendimenti effettivi e quelli attesi sono nulli, il che implica che le informazioni contenute in Φ_t non sono utili per generare rendimenti attesi in eccesso sistematicamente positivi. Se Φ_t contiene informazioni su come ottenere un rendimento in eccesso atteso in t+1, allora quell'informazione è già incorporata nel prezzo P_t e non in P_{t+1}.
  Proprio come asserito in precedenza, non è necessario che il valore atteso dei rendimenti in eccesso sia esclusivamente uguale a zero: l'importante è che tali aspettative non siano distorte.
- Modello Martingala: la nozione di Martingala riproduce il concetto di fair game, ma con questo approccio vengono modellate le proprietà statistiche delle serie temporali dei rendimenti azionari dei prezzi. Per definizione, la Martingala è un processo determinato

rispetto alle osservazioni passate dell'evento che si prende in considerazione. Una variabile statistica segue un processo dinamico Martingala rispetto a Φ, se il suo valore atteso futuro condizionale, data la sequenza di Φ, è uguale al suo valore corrente, ovvero se

$$E[Y_{t+1}|Y_0, \dots, Y_t] = Y_t$$

Poiché l'uso efficiente delle informazioni implica che il futuro prezzo atteso sia uguale al prezzo corrente moltiplicato per (1+rendimento atteso), ovvero

$$E[P_{t+1}|\Phi_t] = P_t(1 + E[r_{t+1}|\Phi_t])$$

allora risulta che

$$P_t = \frac{E[P_{t+1}|\Phi_t]}{(1 + E[r_{t+1}|\Phi_t])}$$

ovvero il prezzo corrente corrisponde al valore attuale del futuro prezzo atteso, scontato al tasso di rendimento atteso congruo per rischio. Dunque il prezzo atteso properly discounted è una Martingala.

- Modello Random Walk: con questo approccio viene definita l'intera forma della distribuzione statistica dei rendimenti in eccesso (gli approcci fair-game e martingala specificano solo il valore atteso e non considerano gli altri momenti della distribuzione). Secondo questo modello, le variazioni successive dei prezzi seguono un processo random walk se sono iid, cioè identically and indipendent distributed, ovvero se hanno la stessa distribuzione di probabilità e se sono statisticamente indipendenti.
  Dunque, i processi random walk hanno la proprietà di essere senza memoria, per cui la conoscenza degli stati passati in cui si è trovato il sistema non è utile per predirne lo stato futuro. Applicando questa proprietà ai prezzi e ai rendimenti azionari, si può perciò affermare che la conoscenza dei prezzi o dei rendimenti passati non è utilizzabile per predire prezzi e rendimenti futuri. In termini statistici, con il random walk, il rendimento atteso è costante ed il rendimento atteso in eccesso ha media nulla. Se il processo di formazione dei prezzi segue un random walk significa che il valore atteso degli excess returns è nullo e che l'unico determinante del valore atteso delle quotazioni azionarie è il valore attuale del flusso di dividendi pagati dalle imprese.
  Questo modello attribuisce alla serie dei prezzi la caratteristica secondo la quale tutte le successive variazioni di prezzo rappresentano scostamenti casuali dai prezzi precedenti, e quest'immagine di casualità ha alimentato la percezione del mercato finanziario come un casinò, in cui l'investimento in titoli è assimilabile al gioco d'azzardo e gli investitori non sono altro che giocatori, i cui profitti e perdite sono legati esclusivamente alla fortuna. Invece, la casualità del movimento dei prezzi è esattamente il comportamento che caratterizza un mercato finanziario che elabora in modo efficiente l'informazione disponibile agli operatori: poiché si può osservare un cambiamento dei prezzi solo nel caso in cui arrivino nuove informazioni, alle quali viene attribuito un significato da parte degli operatori, che viene rivelato dai prezzi di mercato, e poiché il processo di rivelazione di queste nuove informazioni è per definizione casuale, in quanto, se le informazioni fossero

in qualche modo prevedibili, quindi non casuali, esse non sarebbero di fatto nuove, ma anzi sarebbero già anticipate dagli operatori e già incluse nella formazione dei prezzi, allora le variazioni dei prezzi devono essere per forza casuali.

Inoltre, l'ipotesi del random walk non implica il fatto che le variazioni successive dei prezzi siano puramente casuali, senza alcuna ragione economica che supporti questo cambiamento: i prezzi cambiano poiché l'impresa si evolve in continuazione, così come si evolvono la percezione e la valutazione delle caratteristiche dell'impresa da parte degli operatori.

In sintesi, la logica che sta alla base dell'ipotesi del random walk afferma che, dato un certo momento, se il flusso di informazioni non è ostacolato e se le informazioni si riflettono immediatamente nei prezzi delle azioni, allora la variazione dei prezzi nel momento successivo riflette solo le novità del che si manifestano in tale momento ed è casuale ed indipendente rispetto allo stato delle attuali informazioni e rispetto al cambiamento dei prezzi avvenuto nel momento attuale.

Dato che l'informazione è disponibile a tutti, che tutte le novità sono prontamente incorporate nei prezzi e che questi ultimi dimostrano di avere delle variazioni successive che si comportano secondo un processo di random walk, allora la diretta conseguenza dell'EMH si traduce nel fallimento di qualsiasi strategia di investimento che miri a battere il mercato in modo sistematico e nel fatto che nessuno abbia la possibilità di trarre profitto più di qualcun altro. Statisticamente, un investitore può ottenere gli stessi risultati sia allocando la sua ricchezza su titoli a caso, sia seguendo dei consigli professionali, o in altre parole un investitore alle prime armi e un investitore esperto possono raggiungere le stesse performance. Burton Malkiel, nel suo libro "A Random Walk Down Wall Street", afferma che uno scimpanzé bendato potrebbe selezionare, lanciando freccette a caso sul Wall Street Journal, un portafoglio di investimenti tale da ottenere gli stessi risultati di un portafoglio creato da un gruppo di persone esperte; tali risultati sarebbero pari al rendimento medio di mercato. Infatti, l'EMH suggerisce che sarebbe più redditizio per un operatore investire il proprio denaro in un fondo indicizzato, piuttosto che cimentarsi nella costituzione di un portafoglio articolato o subire l'onere degli elevati costi di transazione legati alle numerose operazioni effettuate.

Fama ha sviluppato il concetto di "informazione disponibile" ricorrendo alla seguente tassonomia con contenuto di informazioni progressivamente crescente, con cui ha identificato tre forme di efficienza di mercato:

- Forma debole: l'informazione è limitata alla serie storica dei prezzi azionari passati, che è il dato pubblico più diffuso e facilmente conoscibile dagli operatori. In altre parole, i prezzi riflettono correttamente ed istantaneamente tutte le informazioni presenti nei prezzi passati.

  Se questo tipo di efficienza è verificato, allora non è possibile ottenere rendimenti sistematicamente superiori rispetto alla media di mercato, utilizzando solo l'informazione contenuta nelle serie storiche. Questa tipologia di efficienza nega la validità dell'analisi tecnica, la quale ha l'obiettivo di prevedere le variazioni di prezzo sui mercati finanziari attraverso l'analisi delle statistiche di mercato e dello storico dei grafici che descrivono l'andamento del prezzo. L'analisi tecnica consiste nell'utilizzo di una serie di strumenti che facilitano la comprensione dell'andamento dei grafici e permettono l'identificazione di schemi da parte degli operatori finanziari tecnici.

All'interno di questa forma di efficienza di mercato, rimane la possibilità che un eccesso di rendimento possa essere realizzato utilizzando l'analisi fondamentale.

- Forma semi-forte: l'informazione comprende, oltre alle informazioni racchiuse nei prezzi passati, anche i bilanci, le notizie stampa, i dati macroeconomici, finanziari, settoriali, ovvero l'intero insieme dei dati pubblicamente disponibili.
  Se questo tipo di efficienza è verificato, allora non è possibile ottenere rendimenti sistematicamente superiori rispetto alla media di mercato, utilizzando strategie basate su questo livello di informazione. Ciò implica una negazione della validità sia dell'analisi tecnica che dell'analisi fondamentale, in quanto qualsiasi informazione reperita attraverso uno di questi due metodi sarebbe già disponibile e dunque già incorporata nei prezzi correnti. L'analisi fondamentale si basa sull'identificazione e la previsione di variabili economiche e finanziarie che influenzano l'andamento delle quotazioni delle azioni, tramite l'analisi di due macro-aree di informazioni: la prima riguarda indicatori relativi al sistema economico nel suo complesso, quali ad esempio PIL e tasso d'inflazione, mentre la seconda, che si fonda su un approccio microeconomico, riguarda le informazioni relative alla solidità patrimoniale e alla redditività attesa delle società emittenti, in relazione al prezzo di mercato dei loro titoli all'interno delle prospettive di crescita del settore di appartenenza.
  All'interno di questa forma di efficienza, solo chi detiene informazioni private può avere un vantaggio, in quanto questa tipologia di informazioni non è contenuta nei prezzi.
- Forma forte: l'informazione comprende tutti i dati, anche quelli non pubblicamente disponibili (insider information). Dunque, in tale contesto, i prezzi di mercato rispecchiano qualunque tipo di informazione concernente il titolo finanziario in considerazione.
  Se questo tipo di efficienza è verificato, allora nessun operatore, neanche se fosse in grado di sfruttare informazioni rilevanti e private riguardanti una società in particolare, sarebbe in grado di realizzare profitti costantemente positivi.

L'efficienza in forma forte include l'efficienza semi-forte, che a sua volta include l'efficienza in forma debole.

1.3 EVIDENZE EMPIRICHE A FAVORE DELL'EMH

Molte relazioni e ricerche riguardanti l'efficienza dei mercati si sono svolte con l'intento di dimostrare o confutare le tre tipologie di efficienza, classificate da Eugene Fama, appena esposte.

Poiché l'efficienza in forma debole presuppone che i prezzi riflettano solamente le informazioni derivanti dalla serie storica dei prezzi, allora, per verificare la sussistenza di una tal forma di efficienza, sarebbe opportuno esaminare la connessione esistente tra i rendimenti storici di un titolo. Gli studi condotti con questo obiettivo hanno, dunque, utilizzato la correlazione seriale o autocorrelazione, ovvero l'intensità del legame esistente tra i rendimenti nel tempo di uno stesso titolo, come misura per accertare la presenza dell'efficienza in forma debole. Un coefficiente di correlazione seriale positivo indica una relazione diretta tra i rendimenti storici di un titolo, ovvero rendimenti elevati in un determinato momento preannunciano rendimenti altrettanto soddisfacenti in futuro, mentre un coefficiente di correlazione seriale negativo evidenzia una relazione inversa tra i rendimenti storici di un titolo, quindi rendimenti elevati oggi rappresentano

il presupposto di bassi rendimenti in futuro. Affinché l'efficienza in forma debole sia dimostrata, il coefficiente di autocorrelazione deve essere pari a zero. Un coefficiente nullo è la prova di un'assoluta mancanza di correlazione tra i rendimenti temporali di un titolo.
L'evidenza empirica mostra che la correlazione seriale nei mercati azionari, a meno di decimali dovuti a errori o alla ristrettezza del campione, è nulla. L'assenza di autocorrelazione seriale dei rendimenti deriva dal comportamento random walk dei prezzi, per cui variabili iid implicano assenza di autocorrelazione. Dunque, i prezzi passati non influenzano quelli futuri e non sono quindi utilizzabili per prevedere andamenti in istanti futuri della quotazione. In altre parole la serie storica non è sfruttabile per individuare dinamiche sistematiche che permettano di ottenere rendimenti superiori al mercato.
Il professore francese Bruno Solnik, nel suo articolo del 1973 pubblicato sul Journal of Finance dal titolo "Note on the Validity of the Random Walk for European Stock Prices" ha misurato i coefficienti di correlazione seriale per le variazioni di prezzo giornaliere, settimanali e mensili in nove paesi, arrivando alla conclusione che, avendo rilevato correlazioni estremamente ridotte, come si può notare in figura 1, non è possibile formulare strategie d'investimento redditizie basate solamente sull'osservazione della serie storica dei prezzi.

USA	0.03	UK	0.08
France	-0.01	Italy	-0.02
Germany	0.08	Holland	0.03
Belgium	-0.02	Switzerland	0.01
Sweden	0.06		

Figura 1: coefficienti di autocorrelazione dei rendimenti giornalieri in 9 paesi

Oltre al calcolo statistico dei coefficienti di correlazione seriale, l'autocorrelazione tra i rendimenti può essere rappresentata graficamente tramite la distribuzione dei rendimenti del periodo t-1 rispetto al periodo t. In figura 2 viene mostrata la distribuzione dei rendimenti calcolati sull'indice generale della Borsa Italiana.

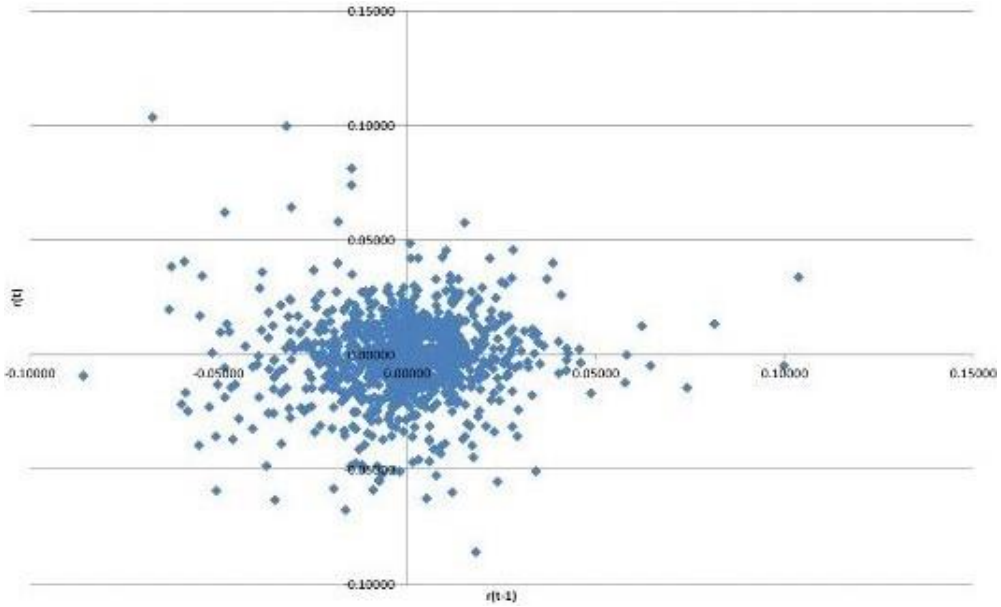

Figura 2: rendimenti logaritmici giornalieri FTSEMIB (gennaio 2008 – gennaio 2021)

Se i rendimenti fossero positivamente correlati, la nuvola dei punti sarebbe concentrata nei quadranti di nord-est e sud-ovest, in quanto rendimenti positivi tenderebbero ad essere seguiti da altri rendimenti positivi e rendimenti negativi tenderebbero ad essere seguiti da altri rendimenti negativi. Se la correlazione fosse negativa, rendimenti positivi sarebbero seguiti da rendimenti negativi e questi ultimi seguiti da rendimenti positivi, con il risultato che la nuvola dei punti sarebbe concentrata nei quadranti nord-ovest e sud-est. Poiché dal grafico in Figura 2 si può notare che risulta impossibile individuare un chiaro pattern, allora si può dedurre che vi è assenza di correlazione.

Inoltre, vi sono stati anche studi che hanno provato a confrontare sequenze puramente casuali con l'ordine con cui si succedono i rendimenti positivi e negativi, come quello di Harry Roberts, il quale, nel suo articolo pubblicato sul Journal of Finance nel 1959 dal titolo "Stock Market "Patterns" and Financial Analysis: Methodological Suggestions", notò come una serie temporale generata da una sequenza di numeri casuali avesse lo stesso aspetto di una serie temporale dei prezzi azionari.

Ma la presenza di autocorrelazione è la logica di fondo dell'analisi tecnica, metodo su cui si basano numerose strategie di investimento, secondo le quali i futuri rendimenti, così come i prezzi, possono essere previsti elaborando le informazioni incorporate nei rendimenti o nei prezzi passati.

La ricerca scientifica e la pratica dell'investimento finanziario si sono dunque cimentate nella determinazione di varie tipologie di strategie apparentemente profittevoli, basate sull'individuazione di pattern di prezzi o sulla verifica di filtri statistici.

Un approccio che utilizza filtri azionari per costruire delle verifiche sulle strategie di investimento fu adottato da Eugene Fama, come riportato nel suo articolo del 1965 dal titolo "The Behavior of Stock Market Price" pubblicato sul Journal of Business, nel tentativo di individuare dei segnali di acquisto o vendita dei titoli. Tale approccio consiste nell'ipotizzare l'esistenza di una fascia di valori intorno ai prezzi azionari, e nell'osservare il movimento dei prezzi rispetto a questa fascia. Se i prezzi rimangono all'interno della fascia, allora si ritiene che essi non comunichino alcuna indicazione utile per decidere sull'investimento azionario e, quindi, si possono trascurare tali variazioni, non acquistando o vendendo il titolo, mentre, se i prezzi sforano le soglie della banda verso l'alto o il basso, allora si ritiene che al mercato siano arrivate nuove informazioni importanti, che il mercato usa per aggiustare il prezzo dell'azione verso una nuova zona di equilibrio. Quindi, se si crede che l'aggiustamento del prezzo verso un nuovo livello stabile sarà graduale, si può utilizzare quel segnale per alimentare la propria strategia d'investimento. Il filtro può anche essere definito su base multiperiodale, per cui si osserva quante volte avviene lo sforamento della fascia di prezzo individuata.

Alcune ricerche econometriche, come quella eseguita da Werner De Bondt e Richard Thaler, economisti noti nel campo della finanza comportamentale, dal titolo "Does the Stock Market Overreact?", pubblicata nel 1985 sul Journal of Fianance, si sono focalizzate sull'individuazione e sulla spiegazione di persistenze nel comportamento dei prezzi azionari, rivelando così l'esistenza di autocorrelazione tra i prezzi. Alcuni di questi comportamenti sono stati denominati momentum ed evidenziano autocorrelazione positiva, in quanto le variazioni dei prezzi seguono l'andamento di quelle precedenti, altri invece sono stati definiti contarian e manifestano autocorrelazione negativa, poiché il comportamento recente dei prezzi volgerà nella direzione opposta in futuro. Questi pattern, che possono essere sfruttati per realizzare extra rendimenti, sono stati spiegati, nell'articolo precedentemente citato, alla luce di una reazione del prezzo alle nuove informazioni diversa rispetto all'aggiustamento istantaneo che rappresenta una delle fondamenta dell'EMH,

come si può notare dalla figura 3. Questa reazione può verificarsi in modo estremamente prudente per quanto riguarda i momentum, per cui si parla di "underreaction", ovvero i prezzi reagiscono lentamente alla notizia e per raggiungere un livello che incorpora l'effetto reale dell'informazione occorrono diversi periodi, oppure può avvenire in maniera eccessiva per quanto riguarda i contrarian, per cui si parla di "overreaction", ovvero i prezzi subiscono una variazione di gran lunga maggiore rispetto a quella che dovrebbe verificarsi, e, come nell'eventualità precedentemente descritta, è necessario un intervallo di tempo relativamente esteso affinché si ritorni ad un valore stabile.

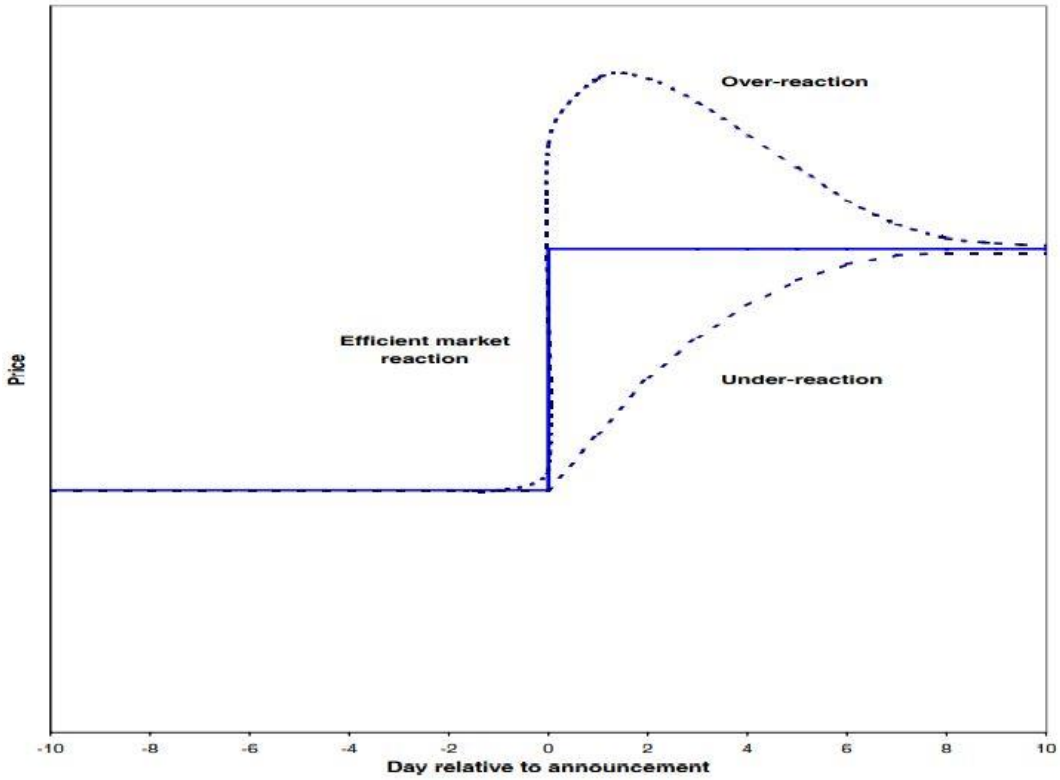

Figura 3: underreaction e overreaction

L'evidenza empirica mostra la presenza di momentum se si prendono in considerazione intervalli di tempo ristretti, mentre nel lungo periodo affiorano contrarian patterns. Tale evidenza è stata verificata dai professori Andrew Lo e Craig MacKinlay, i quali, nel libro intitolato "A Non-Random Walk Down Wall Street" del 1999, calcolando i rendimenti settimanali e mensili per vari indici azionari, hanno trovato prove di una correlazione seriale positiva, il che implica la presenza di momentum nei prezzi delle azioni considerate.

Entrambi i patterns delineati possono derivare dalla difficoltà e dall'incertezza degli operatori ad individuare il nuovo valore fondamentale dell'azione dopo l'arrivo della nuova informazione.

Eugene Fama e il professore statunitense Kenneth French, nel loro articolo del 1988 intitolato "Permanent and Temporary Components of Stock Prices" pubblicato sul Journal of Political Economy, tralasciando l'aspetto relativo all'emotività delle persone che comporta una determinata reazione alle notizie, presentano il fenomeno definito "mean reversion" quale spiegazione del contrarian pattern: i prezzi hanno la tendenza di tornare verso la media nel lungo periodo. Tuttavia, la constatazione del ritorno alla media non è uniforme in tutti gli studi ed è un po' più debole in alcuni periodi rispetto ad altri. Infatti, i risultati empirici che dimostrano, con più determinazione,

la presenza di mean reversion provengono da studi effettuati considerando come campione intervalli temporali che includono la Grande Depressione, periodo in cui i dati possono portare alla generazione di modelli difficilmente generalizzabili.
Inoltre, sebbene questi patterns siano stati identificati come fenomeni utili alla costituzione di strategie di investimento al fine di ottenere profitti superiori alla media, tale possibilità risulta non essere definibile come conseguenza diretta della semplice esistenza dei patterns. A tale proposito, Burton Malkiel, la professoressa Zsuzsanna Fluck e l'economista Richard Quandt, nel loro paper del 1997 intitolato "The Predictability of Stock Returns: A Cross-Sectional Simulation", pubblicato sul Review of Economics and Statistics, hanno simulato una strategia di acquisto di azioni su un periodo di 13 anni, tra gli anni '80 e i primi anni '90. L'acquisto ha riguardato due gruppi distinti di azioni, in base al rendimento che queste mostravano in un periodo di 3-5 anni precedente all'intervallo temporale preso come riferimento per l'attuazione della simulazione. Il gruppo, le cui azioni avevano avuto rendimenti scarsi, tendeva a rivelare rendimenti molto più alti nel periodo successivo, mentre, viceversa, le azioni che avevano avuto rendimenti considerevoli nel periodo precedente mostravano poi rendimenti più ridotti. Pertanto, il loro studio ha confermato l'evidenza statistica di inversioni di rendimento, ma, poiché hanno anche rilevato che, dopo l'inversione di tendenza, i rendimenti per entrambi i gruppi erano simili, non hanno potuto confermare che una strategia basata sull'individuazione di un pattern mean reversion garantisse rendimenti in eccesso agli investitori.
In conclusione, nonostante l'evidenza statistica abbia prodotto risultati misti e contraddittori, trovando anche sequenze persistenti nei movimenti dei titoli e nella variazione di prezzi e rendimenti, nella maggior parte dei casi in cui le strategie, basate sullo sfruttamento di questi patterns, sembrano profittevoli, se si considerano i costi di transazione necessari ad implementarle e se si tiene conto del fatto che, in un mercato efficiente, le strategie di sfruttamento dei patterns individuati sono proprio quei comportamenti che consentono di scoprire le tendenze dei prezzi, costringendoli così ad aggiustamenti verso il valore corretto, allora i profitti risultano fortemente ridimensionati, temporanei o scompaiono o restano circoscritti a determinati periodi o portafogli di titoli per poi svanire. Dunque, in generale l'ipotesi di casualità delle variazioni dei prezzi e dei rendimenti non può essere esclusa anche a livelli severi di significatività.

Per verificare l'ammissibilità dell'efficienza in forma semi-forte, la quale prevede l'assorbimento da parte dei prezzi di tutte le informazioni pubblicamente disponibili, sono stati condotti degli event studies che analizzano il comportamento dei prezzi a seguito di eventi o dell'annuncio di nuove operazioni, anche solo potenziali, da parte delle imprese o della comunicazione di informazioni contabili e finanziarie. Il primo studio in tal senso fu compiuto da Eugene Fama, Lawrence Fischer, Michael Jensen e Richard Roll, i quali, nel 1969, pubblicarono il loro articolo dal titolo "The Adjustment of Stock Prices to New Information" sull'International Economic Review. A seguito dell'evento, se, data la dinamica dei prezzi, i rendimenti anomali restano piatti, allora l'evidenza è coerente con l'ipotesi di efficienza formulata. Sono stati analizzati i seguenti eventi: stock splits e reverse stock splits, ovvero lo scambio del numero di azioni con un altro rispettivamente maggiore o minore, a parità di valore nominale complessivo del capitale sociale; notizie sugli utili (earnings announcements), e sulle "sorprese" degli utili (earnings surprises) rispetto alle aspettative degli analisti finanziari che operano sul mercato; operazioni di fusioni ed acquisizioni (m&a); emissioni azionarie; riduzioni o sospensioni di dividendi.

I risultati delle verifiche empiriche su questi temi sono contrastanti a seconda dei periodi, degli eventi e dei portafogli di titoli presi in considerazione. Le conclusioni nella maggioranza degli studi, di valenza a volte limitata e soggetti a diverse contestazioni di natura metodologica, possono essere sintetizzate come segue:

- Per quanto riguarda la variazione dei prezzi a seguito di operazioni di stock split, poiché si tratta di una procedura puramente formale che non impatta sul valore detenuto dagli azionisti, si è verificato che i prezzi non reagiscono in modo significativo e gli operatori non guadagnano rendimenti in eccesso significativamente diversi da zero, a meno che non vi sia un'aspettativa tale per cui l'operazione di stock split possa determinare un miglioramento della politica di dividendo.
- L'analisi della relazione tra utili e prezzi azionari ha dimostrato che avviene una variazione dei prezzi solamente quando gli utili differiscono da quelli che gli analisti finanziari presenti sul mercato si aspettano, altrimenti si ritiene che tale evento non sia foriero di nuovo contenuto informativo. Nonostante ciò, la maggior parte delle informazioni contenute negli utili è anticipata dal mercato prima del momento del loro annuncio e della comunicazione al pubblico. Infatti, come riportato nell'articolo precedentemente citato, è stato dimostrato che i prezzi azionari iniziano a reagire alle notizie riguardanti gli utili ben dodici mesi prima del momento dell'annuncio e che il movimento dei prezzi prosegue fino a due mesi dopo l'annuncio. Altre ricerche hanno confermato risultati analoghi, estesi anche alle earning surprises. Per sfruttare i post-earnings-announcement-drift, derivati dalla differenza tra gli utili attesi e quelli comunicati, gli arbitraggisti devono investire rilevanti somme in pochi titoli, esponendosi così al rischio specifico di tali titoli, stimabile con la quota della varianza del rendimento non spiegata dal modello di mercato. Perciò, il rendimento in eccesso di un portafoglio che sfrutti i post-earnings-announcement-drift dovrebbe essere spiegato come una logica conseguenza del maggiore rischio assunto. Dunque, tenendo in considerazione il rischio di arbitraggio e i costi di transazione, un investitore non potrebbe ottenere profitti in eccesso significativamente diversi da zero, approfittando delle variazioni dei prezzi a seguito di annunci relativi agli utili.

  Inoltre, è stato verificato che i prezzi delle azioni non cambiano in rapporto ai diversi metodi contabili adottati dalle imprese, in quanto il mercato è in grado di distinguere tra variazioni degli utili indotte da fattori reali e variazioni determinate da modifiche dei principi contabili senza che le caratteristiche operative della gestione giustifichino tale scelta. Di conseguenza, il ricorso a tecniche di "creative accounting", relative alla valutazione di diverse voci contabili, quali ad esempio ammortamento, scorte, e partecipazioni, effettuato con il solo scopo di migliorare l'immagine aziendale, si rivela insignificante ai fini della valutazione di mercato. La stessa conclusione può essere adottata in relazione alle tecniche di "accounting conservatism", secondo le quali vengono applicate delle linee guida contabili caratterizzate da un elevato grado di prudenza, che consistono nell'anticipare le perdite non appena sene viene a conoscenza e nel rilevare i guadagni solo quando sono realizzati.
- A seguito delle comunicazioni di operazioni di fusione o di acquisizione, sono stati individuati post-event-drift, il cui sfruttamento dipende però dalle modalità con cui viene pagata l'operazione: se regolata per cassa, le opportunità di sfruttamento economico dei drift sono irrilevanti, se regolata con emissioni azionarie, i drift conducono a significative perdite.

- In merito ad eventi quali emissione di nuove azioni o IPO, sono state rilevate evidenze contradditorie.
- Sono state rilevate significative reazioni negative all'annuncio della riduzione o del taglio dei dividendi, con una durata fino a 12 mesi dopo il momento della comunicazione. Tuttavia, poiché, nella maggior parte dei casi, le politiche di dividendo subiscono dei cambiamenti a causa di variazioni importanti degli utili, se si tiene conto della performance degli utili dell'impresa prima dell'evento, allora le reazioni dei prezzi tendono a sparire, ovvero tendono ad essere spiegate dagli earnings announcements piuttosto che dai dividend announcements.

Dunque, in conclusione, lo studio dell'impatto di eventi aziendali sui prezzi azionari accetta l'esistenza di significative e persistenti opportunità di profitti in eccesso, ma, considerando le difficoltà cui si andrebbe incontro per cercare di sfruttarle, l'ipotesi di mercato efficiente non dovrebbe essere abbandonata.

Un altro metodo per convalidare l'efficienza in forma semi-forte consiste nell'analisi dei rendimenti dei fondi comuni d'investimento, in quanto gli investitori professionali, che lavorano in questi fondi, basano le loro valutazioni sulle informazioni pubbliche, oltre che sulle serie storiche dei prezzi. Inoltre, poiché questi investitori professionali sono spesso compensati tramite incentivi, nel caso riescano ad ottenere rendimenti superiori alla media, allora questa analisi risulta cruciale nella verifica dell'efficienza del mercato. L'analisi delle performance dei fondi di investimento è stata esaminata a lungo e sotto diverse prospettive, ma, in linea generale, si nota che i loro rendimenti sono simili a quelli del mercato e che non esiste nessun investitore professionale in grado di battere sistematicamente il mercato.

Uno dei primi studi sulla performance dei fondi comuni d'investimento è stato condotto da Michael Jensen, il quale ha pubblicato, nel 1968, sul Journal of Finance, l'articolo "The Performance of Mutual Funds in the Period 1945-1964". Nella valutazione effettuata, Jansen ha utilizzato il β del CAPM (Capital Asset Pricing Model) come misura del rischio, per aggiustare i rendimenti osservati per il rischio sistemico implicito in essi, e ha scoperto che i fondi comuni d'investimento gestiti attivamente tendevano a sottoperformare il mercato in misura pari alle spese realizzate per la gestione del fondo stesso.

Nel 1997, sul Journal of Finance, Mark Carhart ha pubblicato il suo articolo intitolato "On Persistence in Mutual Fund Performance", in cui ha utilizzato un diverso metodo di aggiustamento del rischio, nella valutazione della performance dei fondi comuni di investimento a gestione attiva. Il rischio è stato misurato con un modello a quattro fattori, in cui, oltre al β, vengono tenuti in considerazione anche il fattore riguardante la dimensione dell'impresa, secondo cui un'impresa di piccole dimensioni deve essere associata ad un rischio maggiore, il fattore "value", per cui aziende con un rapporto P/E ridotto comportano un rischio rilevante, e il fattore momentum. Nello studio condotto, è stato riscontrato che i rendimenti ottenuti, al netto del rischio e delle spese di gestione, non erano maggiori rispetto a quelli ottenuti da un fondo indicizzato.

Le analisi compiute sui rendimenti dei fondi comuni d'investimento devono tenere conto di alcuni bias significativi, come ad esempio il "survivorship bias", a causa del quale i dati su cui vengono effettuati gli studi possono risultare distorti. Questo bias si riferisce alla situazione per cui molti fondi con prestazioni scadenti vengono inglobati da altri fondi, distruggendo così i dati negativi dei fondi che, da quel momento in poi, non esistono più come identità unica, e mostrando solamente i rendimenti generati dai fondi in vita. Burton Malkiel, nel suo paper del 2011, intitolato "The Efficient-Market Hypothesis and the Financial Crisis", evidenzia il fatto che i rendimenti ottenuti dai

fondi esistenti, ovvero quelli sopravvissuti, sono notevolmente migliori se comparati con quelli effettivi di tutti i fondi, compresi quelli liquidati o fusi in altri fondi comuni di investimento, come si può vedere in figura 4.

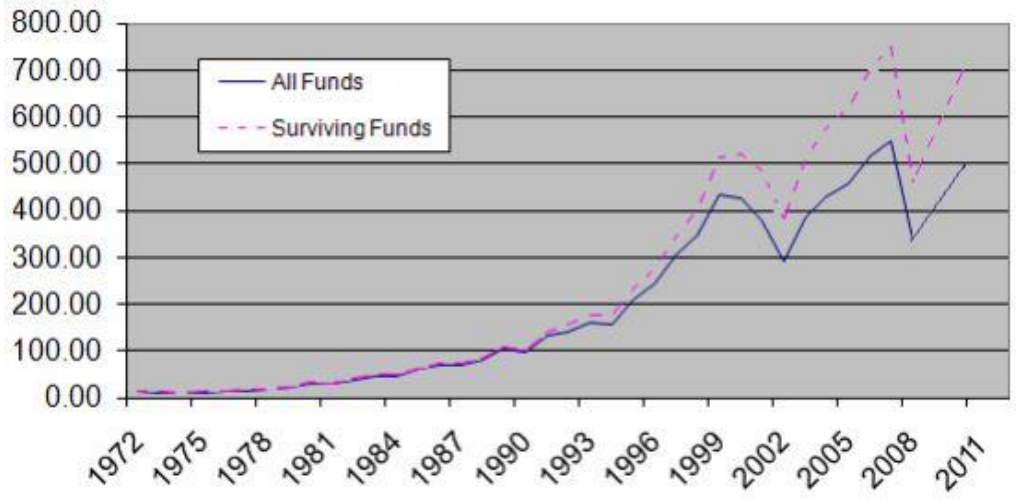

Figura 4: rendimenti dei fondi comuni d'investimento sopravvissuti vs rendimenti di tutti i fondi comuni d'investimento

Nonostante la distorsione dei rendimenti mostrati, Malkiel afferma che gli investitori professionali non riescono a sovraperformare il mercato, anche se nelle analisi vengono utilizzati dati con un certo grado di "survivorship bias". Infatti, come viene mostrato in figura 5, secondo lo studio condotto da Malkiel, compiuto sui fondi statunitensi tra il 2005 e il 2010, circa due terzi dei fondi a gestione attiva ottengono risultati minori di quelli dei relativi indici presi come riferimento. Egli ha operato una suddivisione dei fondi comuni d'investimento in base alla capitalizzazione delle aziende su cui vertono gli investimenti e in base alla diversificazione geografica, andando a confrontare tali fondi con i rispettivi indici della serie S&P ponderati sulla base delle stesse caratteristiche evidenziate per i fondi comuni d'investimento.

Fund Category	Benchmark Index	Percent Outperformed
All Domestic Equity	S&P 1500	57%
All Large Cap Funds	S&P 500	62
All Mid-Cap Funds	S&P Mid-Cap 400	78
All Small-Cap Funds	S&P Small-Cap 600	63
All Multi-Cap Funds	S&P Small-Cap 1500	66
Global Funds	S&P Global 1200	59
International Funds	S&P 700	85
Emerging Market Funds	S&P/IFCI Composite	86

Figura 5: percentuale dei fondi sovraperformati dal relativo indice benchmark

Sicuramente, si possono trovare dei fondi comuni di investimento che generano rendimenti di gran lunga superiori alla media all'interno di un determinato arco di tempo, ma questi risultati si rivelano poco persistenti, in quanto non si verificano nel periodo successivo. Malkiel ha evidenziato il fatto che, durante gli anni Settanta, i 20 fondi comuni d'investimento più grandi negli Stati Uniti abbiano

ottenuto rendimenti quasi doppi rispetto alla performance del benchmark, ma che nella decade successiva i risultati si siano rovesciati, a favore di una migliore performance dell'indice di riferimento, e che la stessa situazione si sia ripetuta per i decenni successivi.
Tramite un'analisi durata quarant'anni, tra il 1970 e il 2010, Malkiel ha studiato l'andamento del numero dei fondi comuni d'investimento degli Stati Uniti e dei rendimenti da questi ottenuti confrontandoli con l'indice S&P500, benchmark spesso utilizzato per misurare i rendimenti complessivi del mercato azionario. Dei 353 fondi comuni d'investimento esistenti nel 1970, solamente 108 di essi sono rimasti attivi fino al 2010. Premettendo che i 245 fondi comuni d'investimento che sono stati incorporati in altri fondi avranno sicuramente avuto performance peggiori rispetto a quelle dimostrate dai fondi ancora in vita, è comunque impossibile ammettere che questi investitori professionali riescano mediamente a battere il mercato. Infatti, come si può notare dalla figura 6, il 61% dei fondi ancora esistenti nel 2010, ovvero 66 fondi su 108, hanno avuto rendimenti annualizzati minori rispetto al benchmark per tutta la durata dell'analisi. Solamente cinque fondi comuni d'investimento sono riusciti ad ottenere performance maggiori, tra i due e i quattro punti percentuali in più, rispetto all'indice S&P500.

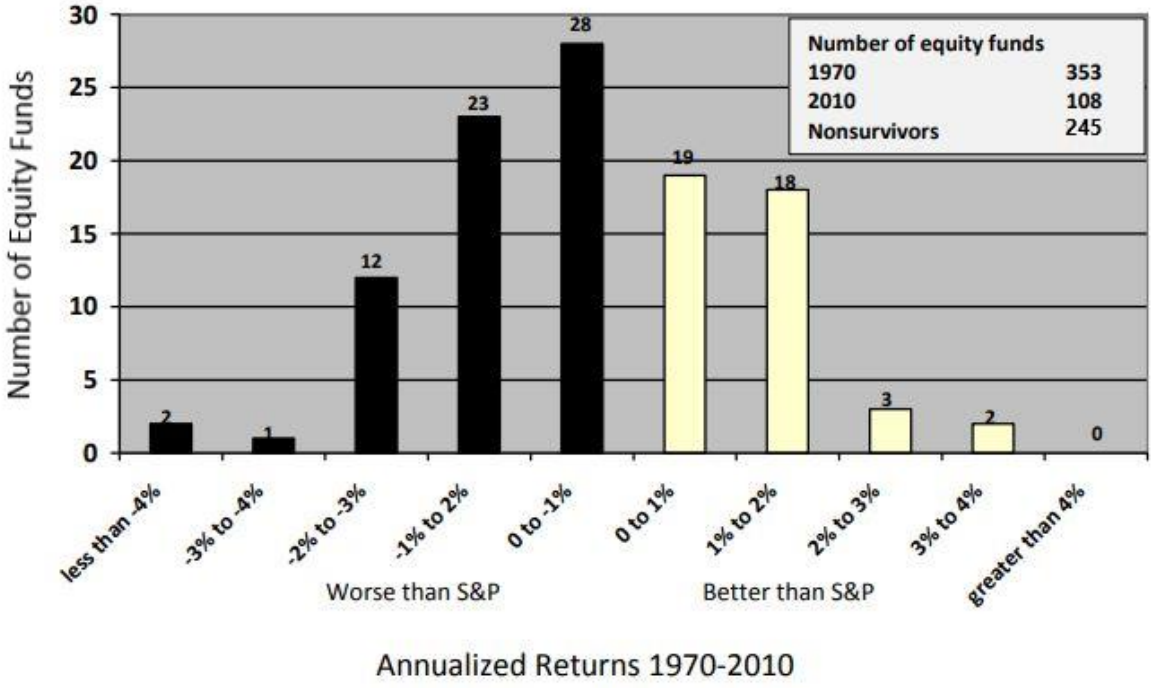

Figura 6: rendimenti dei fondi comuni d'investimento sopravvissuti tra il 1970 e il 2010 vs rendimenti S&P500

I sostenitori dell'analisi fondamentale, quale metodo su cui basare le proprie scelte di investimento, apportano, come esempi riguardanti l'inammissibilità dell'efficienza del mercato in forma semi-forte, evidenze di investitori professionali, quali ad esempio Warren Buffet, che riescono a battere ripetutamente il mercato. L'oracolo di Ohama, così come è conosciuto uno dei più grandi value investors di sempre, afferma "I'd be a bum in the street with a tin cup if the markets were efficient". Però, come analizzato dagli studi citati in precedenza, si può dimostrare che queste eventualità non sono la prassi nel mondo degli investimenti, ma possono essere considerate come eventi estremi, molti distanti dalla media delle osservazioni effettuate. Le persistenze di performance migliori della media di mercato appaiono temporanee e decadono all'aumentare dell'arco di tempo su cui sono misurate.

Dunque, l'evidenza empirica ha prodotto risultati misti e contrastanti, ma in generale ha messo in luce che la maggioranza dei fondi di investimento non sono in grado di battere sistematicamente le semplici strategie di buy-and-hold o di fare meglio dell'indice generale di borsa.
L'interpretazione assegnata da molti economisti è che tali risultati tendono a confermare l'ipotesi di efficienza del mercato in forma semi-forte.

Per quanto riguarda invece l'efficienza in forma forte, questa tipologia ha avuto un processo di verifica complicato, in quanto la presenza di informazioni riservate comporta una situazione di rilevanti differenze nella quantità e qualità dell'informazione tra i vari operatori, minando così le conclusioni dell'economia finanziaria neoclassica e rendendo non ottimale l'allocazione delle risorse che deriva dal processo di ricerca dell'equilibrio di mercato.
L'analisi delle reazioni dei prezzi azionari alle informazioni riservate, è stata più complessa da approfondire a causa delle notevoli limitazioni nella disponibilità di dati. Gran parte delle ricerche in materia hanno fatto ricorso alle indagini sul reato di insider trading, ricavando, dai dossier accumulati nelle Autorità di controllo dei mercati azionari, le informazioni riservate, il timing della loro rivelazione pubblica ed il comportamento dei prezzi prima e dopo la comunicazione al mercato.
La maggior parte delle verifiche statistiche porta alla conclusione che il mercato finanziario non è efficiente nella forma forte: vi sono operatori che, approfittando delle insider information, riescono a prevedere il futuro comportamento dei prezzi ed a sfruttare opportunità di significativi profitti; la maggioranza degli operatori del mercato resta sorpreso dalle informazioni riservate quando queste vengono rese pubbliche, con il risultato che i prezzi si aggiustano impiegando un periodo anche lungo per raggiungere un nuovo livello di equilibrio.
È stato, dunque, dimostrato che l'informazione privata arriva ad essere incorporata nel prezzo solo dopo aver concesso già qualche occasione di profitto agli insiders.
Negli ultimi anni, le norme relative al reato di insider trading stanno diventando sempre più severe e stringenti e i controlli da parte delle autorità addette sempre più minuziosi ed intransigenti. Di conseguenza, gli extra profitti, generati da coloro che basano le proprie operazioni su informazioni riservate, sono sempre meno frequenti.

In sintesi, si può concludere che il mercato finanziario è complessivamente efficiente nella forma debole, è sufficientemente efficiente nella forma semi-forte, con significative eccezioni, ma non è efficiente nella forma forte.

1.4 PUZZLES

Alcuni studi escludono l'ipotesi di efficienza dei mercati finanziari, ritenendo che sia possibile conseguire extra profitti, sfruttando determinati fattori che sarebbero rivelatori dei rendimenti futuri dei titoli. Questi fattori generano delle vere e proprie inefficienze, definite "puzzles", che possono essere persistenti, specifiche di determinati periodi o destinate a scomparire.
Nel corso della storia i puzzles più riscontrati sono:

- Small-firm effect: la dimensione delle imprese è risultata essere un fattore su cui gli investitori hanno messo a punto delle strategie, in quanto, tra il 1926 e il 1975, è stato rilevato che le azioni delle imprese di dimensione minore abbiano realizzato rendimenti

maggiori rispetto a quelli delle imprese più grandi, e che, inoltre, questi rendimenti siano stati superiori rispetto a quelli previsti dal CAPM.
Eugene Fama e Kenneth French, nel 1992, nel loro articolo, pubblicato sul Journal of Finance, dal titolo "The Cross-Section of Expected Stock Returns", hanno esaminato i rendimenti mensili delle azioni statunitensi tra il 1963 e il 1990, dividendo le aziende in decili in base alla loro dimensione, misurata tramite la capitalizzazione totale. Come si può notare dalla figura 7, è presente un'evidente tendenza, che indica rendimenti mensili maggiori per le aziende con capitalizzazione inferiore.

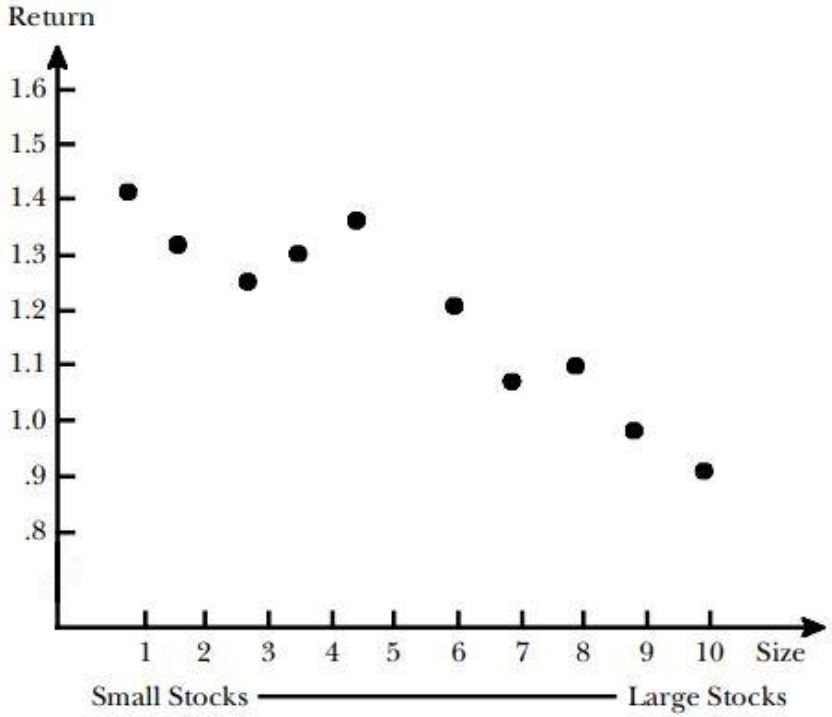

Figura 7: rendimenti mensili tra il 1963 e il 1990 di portafogli creati sulla base della dimensione aziendale

I rendimenti più elevati generati da tali imprese potrebbero, però, essere imputabili ad un rischio maggiore connesso ad un investimento in titoli di imprese con dimensioni ridotte. Sebbene il CAPM assuma che il rendimento di un titolo sia strettamente dipendente dal suo rischio sistematico, incluso nel parametro β, secondo alcuni economisti il rendimento elevato, dimostrato dalle imprese piccole, nel periodo tra il 1936 e il 1990, dovrebbe tenere in considerazione anche il premio per il rischio connesso alla dimensione dell'impresa, oltre a quello per il rischio sistematico. Eugene Fama e Kenneth French suggeriscono che la dimensione possa essere una proxy di rischio migliore rispetto a β, e che, quindi, i loro risultati non debbano essere interpretati come un'indicazione dell'inefficienza dei mercati. L'affidabilità di questo fenomeno, però, è stata messa in discussione, in quanto, dagli studi effettuati tra il 1985 e l'inizio del XXI secolo, non è stato osservato alcun guadagno dalla detenzione di azioni di aziende più piccole; anzi, nella maggior parte dei mercati mondiali, le azioni delle aziende a capitalizzazione maggiore hanno prodotto tassi di rendimento più elevati. È possibile che la crescente istituzionalizzazione del mercato abbia portato gli operatori a preferire le società più grandi con maggiore liquidità, in quanto quelle più piccole potrebbero riscontrare ingenti difficoltà nel liquidare importi significativi di azioni.

Infine, è anche possibile che alcuni studi sull'effetto in questione siano stati influenzati dal "survivorship bias": poiché le aziende a bassa capitalizzazione possono incorrere in un fallimento con più probabilità rispetto a quelle a larga capitalizzazione, ricerche condotte in un determinato momento, inerenti la storia passata delle aziende di piccole dimensioni, misurerebbero solamente le prestazioni delle aziende ancora in attività, tralasciando le performance scadenti di quelle che hanno fallito.

- Calendar effect: questo tipo di inefficienza fa riferimento a delle anomalie temporali, che sono state osservate per diversi anni tra i mesi di dicembre e gennaio e tra i giorni di venerdì e lunedì.

  La prima è nota come turn-of-the-year effect o January effect, e riguarda il fatto che i rendimenti osservati a gennaio sono più elevati rispetto a quelli degli altri mesi. Questo puzzle, che comporta un divario tra i rendimenti, potrebbe essere collegato agli effetti fiscali connessi alla deducibilità delle perdite su titoli, che rendono conveniente la vendita delle azioni che hanno avuto una performance negativa entro la fine dell'anno, per poi comprarle nuovamente all'inizio di gennaio dell'anno successivo: questo meccanismo, che conferisce a questa inefficienza anche il nome di "tax-loss-harvesting effect", porterebbe ad un aumento dei prezzi nel mese di gennaio. Tale puzzle sembra persistere ancora oggi, anche se in modo molto meno evidente rispetto al passato, in quanto una volta pubblicato in letteratura dagli economisti Robert Haugen e Josef Lakonishok con il libro "The Incredible January Effect: The Stock Market's Unsolved Mistery" nel 1987, molti investitori hanno iniziato a sfruttare quest'inefficienza, causando una riduzione del suo effetto sui prezzi.

  La seconda anomalia temporale è nota come week-end effect o day-of-the-week effect, e riguarda il fatto che i rendimenti del venerdì sembrano più elevati, mentre quelli del lunedì sembrano più bassi rispetto a quelli degli altri giorni. Questo puzzle è stato osservato per vari periodi tra il 1885 e il 1977, ma non nel periodo dal 1978 al 2002.

  Al contrario di quanto si possa pensare, l'esistenza di queste inefficienze non è incoerente con il concetto di efficienza dei mercati, poiché l'individuazione e la pubblicazione di fenomeni di tale portata comportano una ricerca assidua da parte degli operatori nel tentare di sfruttarli e, dunque, la pressione competitiva ne farebbe sparire l'effetto.
- Value effect: questo tipo di inefficienza fa riferimento al fatto che, fino al 1991, i portafogli di azioni con basso price/earnings, ovvero contenenti le cosiddette value stock, hanno generano rendimenti maggiori di portafogli di azioni con alti price/earnings, caratteristica delle growth stock. Questo effetto ha ricevuto la sua più famosa approvazione accademica nell'articolo già citato in precedenza "The Cross-Section of Expected Stock Returns", di Eugene Fama e Kenneth French, che hanno definito le azioni value come quelle azioni che scambiano a un prezzo basso rispetto al loro valore contabile.
- Dogs of Dow effect: il rapporto dividend yield/price è stato a lungo considerato un buon indicatore utile a prevedere i rendimenti futuri, in quanto, fino al 1991, gli investitori che hanno acquistato, all'inizio di ogni anno, un paniere di titoli di mercato con rapporto iniziale dividend yield/price alto hanno guadagnato un tasso di rendimento più elevato rispetto a chi ha investito in un portafoglio con basso rapporto dividend yield/price. Questo puzzle è stato pubblicato in letteratura nel 1991 con il libro "Beating the Dow" dell'economista Michael O'Higgins. Il nome di tale puzzle deriva dalla strategia ideata da O'Higgins, secondo la quale comprando, all'inizio dell'anno, i dieci titoli azionari dell'indice Dow Jones Industrial Average con il rapporto dividend yield/price più alto e tenendoli in portafoglio per un anno,

gli investitori dovrebbero ottenere un rendimento maggiore rispetto a quello conseguito dal DJIA, uno degli indici più antichi e spesso considerato come riferimento per la performance del mercato in generale.

- Momentum effect: questo effetto si riferisce al fatto che variazioni dei prezzi azionari tendono ad essere seguite da variazioni dello stesso segno, e sembra essere persistente e statisticamente robusto.

La persistenza dei puzzles può essere interpretata come prova di inefficienza del mercato in determinati periodi e circostanze, oppure come un fenomeno in attesa di spiegazioni razionali o di rilevazione di errori metodologici o carenze statistiche.
La scomparsa di puzzles osservati in passato, invece, sembra essere connessa al fatto che man mano che vengono individuati, i puzzles vengono sfruttati dagli arbitraggisti per ricavarne profitti, fino ad eliminarli dal mercato.
Inoltre, a seguito dell'individuazione di questi puzzles, alcuni studiosi si sono interrogati riguardo l'accuratezza della misurazione del rischio dei titoli azionari espressa dal CAPM tramite il parametro β. Una constatazione anomala relativa ai rendimenti in eccesso, che possono essere ottenuti sfruttando tali inefficienze, si rivela un test congiunto tra l'EMH e le procedure di aggiustamento per il rischio impiegate: se il β è una misura inadeguata del rischio, allora risulta inappropriato considerare i rendimenti in eccesso corretti per il β come incoerenti con le ipotesi di efficienza del mercato. Infatti, nel 1993, Eugene Fama e Kenneth French, nel loro articolo, pubblicato sul Journal of Financial Economics, dal titolo "Common Risk Factors in the Returns on Stock and Bonds", proposero un modello di rischio a tre fattori, il quale sosteneva che il rendimento di un portafoglio dovesse essere spiegato alla luce della dimensione e del rapporto price/earnings delle aziende su cui verte l'investimento, oltre che del rischio sistematico insisto nelle azioni che lo compongono.
In alcuni modelli, viene aggiunto anche il fattore momentum al modello di rischio a tre fattori di Fama e French.

1.5 RAZIONALITÁ LIMITATA E BOLLE SPECULATIVE

L'EMH, oltre che affermare l'assoluta imprevedibilità dei movimenti del mercato e la rigorosa riflessione del set informativo nel livello dei prezzi, implica un comportamento degli operatori efficiente e razionale, così come chi progetta le uscite di sicurezza dei locali pubblici presume che in caso di incendio le persone si metteranno in fila ad aspettare il loro turno per sottrarsi alle fiamme. In realtà, i mercati finanziari, come tutte le organizzazioni istituite dall'uomo, sebbene trovino fondamento sulla massimizzazione di una funzione di utilità, rimangono comunque governati da un sistema di scelte e decisioni prese da esseri umani, e pertanto suscettibili di essere influenzate da variabili sociali, emotive e psicologiche. Infatti, oltre agli studi incentrati sulla ricerca di fattori che possano rivelare i rendimenti futuri, dando coì origine a delle inefficienze, vi sono, in letteratura, molte analisi, che basano i loro risultati sulla visione di un mercato caratterizzato da incompletezze ed imperfezioni informative, in cui gli operatori assumono decisioni in maniera non completamente razionale, o del tutto irrazionale, comportando così una distorsione nell'efficiente allocazione delle risorse.
Questa visione si collega, in termini di conseguenze teoriche, al paradosso formulato dagli economisti Sanford Grossman e Joseph Stiglitz, i quali, nel 1980, pubblicarono, sull'American Economic Review, un articolo intitolato "On the Impossibility of Informationally Efficient Markets".

Il paradosso di Grossman e Stiglitz sostiene che, se si suppone che il mercato sia efficiente, nel senso che i prezzi incorporano le informazioni disponibili, e che la produzione e l'analisi di tali informazioni siano collegate al sostenimento di costi, allora nessuno sarebbe disposto ad addossarsi quei costi, in quanto sarebbe sufficiente, adottando un comportamento da free rider, osservare i prezzi per venire a conoscenza di ciò che è importante conoscere, senza dover sostenere alcun costo. Ma se nessuno ha convenienza a spendere risorse per produrre le informazioni che sono necessarie, allora quelle informazioni non saranno prodotte, non verranno rese disponibili ed il mercato non potrà incorporarle nei prezzi. Come può, dunque, esistere un mercato efficiente dal punto di vista informativo?

Dal punto di vista dell'investitore, il valore dell'informazione corrisponde all'incremento dell'utilità attesa che ottiene riallocando la propria ricchezza alla luce delle nuove informazioni, al netto dei costi sostenuti per produrle. Ma se alcuni investitori, facendo affidamento sul concetto di mercato efficiente, che, essendo caratterizzato da assenza di asimmetrie informative, rende le informazioni di dominio pubblico, decidessero di basare le proprie decisioni sulle informazioni prodotte da altri operatori, potrebbero allora ottenere un guadagno maggiore. In un contesto del genere, quindi, nessuno vorrebbe compiere lo sforzo di produrre informazioni, in quanto non otterrebbe un compenso da questa attività maggiore rispetto a coloro che si comportano da free rider. In conclusione, Grossman e Stiglitz affermano che, in presenza di informazioni costose, non può sussistere un equilibrio in un mercato efficiente. Tale conclusione viene perfettamente sintetizzata da una citazione dello storico della finanza Peter Bernstein, tratta dal suo libro del 1998 intitolato "Against the Gods: The Remarkable Story of Risk": "*The information you have is not the information you want. The information you want is not the information you need. The information you need is not the information you can obtain. The information you can obtain costs more than you want to pay."*

Per ovviare a tale paradosso, nel 1986, l'economista Fischer Black, nel suo paper intitolato "Noise" e pubblicato sul Journal of Finance, ha proposto l'introduzione di una sorgente di rumore, di disturbo, detta appunto noise, che impedisce una completa trasparenza delle informazioni, in modo da assicurare un vantaggio a chi ha sostenuto gli oneri di raccogliere ed elaborare le nuove informazioni al fine di renderle poi disponibili. Con questo accorgimento si può raggiungere un equilibrio imperfetto (noisy rational expectations equilibria), con un mercato caratterizzato da prezzi in parte opachi, che non rivelano tutta l'informazione disponibile, e con due tipi di operatori, definiti da Black come "noise traders", ovvero colore che basano le loro decisioni sul rumore, e "information traders", ovvero gli operatori che sostengono un costo per ottenere delle informazioni su cui sperano di ottenere un guadagno.

Il rumore rende possibile gli scambi continui sui mercati finanziari, in quanto, se non ci fossero i noise traders, la maggior parte degli operatori preferirebbe investire sugli indici generali di borsa, tenendoli in portafoglio e scambiandoli raramente. Dunque, poiché il rumore garantisce la liquidità del mercato, permette agli operatori di osservare i prezzi. Ma, a questo punto, i prezzi rifletteranno sia le informazioni rilevanti pervenute dagli information traders, sia il rumore causato dalle operazioni dei noise traders. Perciò, da un lato, il rumore rende i mercati inefficienti, nel senso che i prezzi non incorporano solo le informazioni disponibili ma assorbono anche una fonte di disturbo, dall'altro lato, preserva la conseguenza dell'ipotesi di efficienza dei mercati, in quanto impedisce di sfruttare le inefficienze, proprio perché le informazioni disponibili risultano distorte a causa del rumore stesso.

Il noise viene identificato come un gran numero di piccoli eventi, che può portare a delle conseguenze molto più impattanti rispetto a quanto possa fare un esiguo numero di grandi eventi. Le cause da cui può scaturire questo disturbo possono essere molteplici: la presenza di operatori che, per esigenze finanziarie personali, come la necessità di liquidare rapidamente le proprie posizioni per assolvere ad altri impegni, interferiscono con il sistema dei prezzi; un processo erratico di offerta di titoli; il comportamento di operatori che, sbagliando ad interpretare l'informazione, agiscono in modo errato sul mercato, oppure effettuano transazioni tralasciando la razionalità su cui si dovrebbero basare le valutazioni economiche, e affidando le proprie decisioni ad altre pulsioni, come il gusto per il rischio, abitudini mentali radicate, o debolezze psicologiche come l'affezione o l'avversione per certe imprese. Anche gli studi di finanza comportamentale (behavioural finance), che esaminano gli effetti della psicologia dell'investitore sui prezzi delle azioni, rivelano che gli investitori sono soggetti a molti pregiudizi come l'avversione per le perdite e l'eccesso di fiducia.

Una delle conseguenze dell'assunzione di queste ipotesi teoriche riguarda il fatto che diventa conveniente adottare strategie di investimento non limitate al semplice buy-and-hold, tipiche dell'assunto dei mercati efficienti, ma estese a sfruttare gli effetti sui prezzi derivanti dalle operazioni effettuate dai noise traders. Tuttavia, il ricorso a strategie basate sull'irrazionalità dei noise traders ha come unica conseguenza quella di propagare il rumore nel mercato finanziario, con potenziali conseguenze destabilizzanti sul sistema dei prezzi.

In generale, il rumore rende molto difficile verificare le teorie pratiche o accademiche sul funzionamento dei mercati finanziari o economici, e, come afferma Black, costringe gli operatori ad agire in gran parte al buio.

L'ipotesi di razionalità limitata, che guida il comportamento effettivo degli investitori, è coerente anche con la visione keynesiana del mercato finanziario. L'economista inglese John Maynard Keynes, nel libro "The General Theory of Employment, Interest and Money" del 1936, ha paragonato la borsa ad un concorso di bellezza, in cui gli operatori, sopraffatti dagli "animal spirits", usano le loro risorse non per comprendere i fattori fondamentali che determinano la performance dell'impresa, ma per prevedere il futuro comportamento degli altri operatori ed immaginare quale possa essere la distribuzione delle loro aspettative soggettive. Di conseguenza, il futuro andamento dei prezzi non è determinato dai fondamentali dell'azienda, ma dalle previsioni che gli investitori formulano sull'operato degli altri investitori.

Inoltre, la razionalità limitata degli operatori è causata anche dall'incertezza che caratterizza gli eventi futuri, la quale non permette agli operatori di effettuare un calcolo perfettamente razionale, basato sui fattori economici, né di mantenere una prospettiva di lungo periodo.

La razionalità limitata è un concetto su cui si basano le maggiori critiche all'ipotesi di efficienza dei mercati, che focalizzano la loro attenzione su eventi eclatanti, quali ad esempio il crash del mercato statunitense del 1987 e la cosiddetta bolla di Internet, in cui i livelli dei prezzi si distaccano in maniera vertiginosa dalle valutazioni fondamentali, incorporando delle prospettive prive di razionalità e buonsenso, che trovano spiegazione soltanto in una moltitudine di fattori, ognuno dei quali, considerato singolarmente, non rappresenterebbe un fattore causale rilevante. L'economista Robert Shiller, nel suo libro del 2000 intitolato "Irrational Exuberance", ha utilizzato il termine "self-fulfilling psychology" per descrivere il fenomeno in cui una serie di situazioni producono un effetto maggiore della somma delle parti.

Nell'ottobre del 1987 si è verificato un crollo inaspettato di 22 punti percentuali del mercato azionario statunitense. La finanza comportamentale spiega il fenomeno basandosi su

considerazioni psicologiche, dal momento che i valori fondamentali non sembra fossero cambiati in quel periodo. Sebbene non si possa escludere l'esistenza di fattori comportamentali e psicologici che hanno alimentato il crash dei mercati, non si può, tuttavia, non effettuare delle considerazioni logiche che possono contribuire a spiegare un così brusco cambiamento delle valutazioni di mercato. Vi sono una serie di fattori che potrebbero aver razionalmente cambiato il punto di vista degli investitori sul corretto valore del mercato azionario nell'ottobre 1987: i rendimenti dei titoli del Tesoro a lungo termine sono aumentati da circa il 9% a quasi il 10,5% nei due mesi precedenti la metà di ottobre; l'allora Segretario del Tesoro, James Baker, aveva avanzato, all'inizio del mese di ottobre, l'ipotesi di un ulteriore deprezzamento del dollaro, aumentando i rischi per gli investitori stranieri e spaventando anche gli investitori nazionali; il Congresso, sempre ad inizio mese, aveva proposto di imporre una tassa sulle fusioni aziendali, che avrebbe reso proibitiva tale attività e che avrebbe potuto porre fine all'ondata di fusioni avvenuta nel periodo precedente. Il rischio che l'attività di fusione potesse essere ridotta aumentava i rischi in tutto il mercato azionario, indebolendo la disciplina sulla gestione aziendale che le potenziali acquisizioni generalmente forniscono. Ciò non significa che i fattori psicologici siano irrilevanti nello spiegare la vertiginosa discesa dei prezzi avvenuta, poiché, senza dubbio, hanno avuto un ruolo rilevante, ma sarebbe un errore ignorare il significativo cambiamento della situazione economica generale di contorno, in cui una serie di eventi "fondamentali" sfavorevoli può aver avuto un effetto cumulativo tale da modificare la percezione del rischio da parte degli operatori. L'economista statunitense Merton Miller, nel suo libro "Financial Innovations and Market Volatility" del 1991, afferma che vi sono stati alcuni eventi esterni che hanno segnalato un possibile cambiamento in quello che fino ad allora era stato un clima politico ed economico molto favorevole per le azioni, e che dunque molti investitori hanno iniziato, contemporaneamente, a ritenere di detenere una quota troppo grande della loro ricchezza in azioni con un alto livello di rischio.

Nel 1994, con la quotazione di Netscape, la società che sviluppò il primo browser commerciale per internet, iniziò un ciclo definito come "New Economy", che determinò, in pochi anni, un sorprendente sviluppo di aziende operanti nel settore Internet, o più generalmente in quello informatico. Viene infatti usato il termine Dot-com per identificare quelle aziende che, nate a seguito del notevole surplus di fondi generati dal venture capital e del grande ottimismo del mercato azionario durante la fine del ventesimo secolo, impostarono un business improntato principalmente all'erogazione di servizi via web. Molte di queste società erano prive di piani aziendali o addirittura di qualsiasi forma di profitto. Quando, nei primi mesi del 2000, la bolla scoppiò, le quotazioni dei titoli del settore persero il 90% o più del loro valore e parecchie società fallirono. Diversamente da quanto accaduto durante il 1987, in questo caso la generazione della bolla speculativa è totalmente da imputarsi all'esuberanza irrazionale che ha travolto gli operatori, ma soprattutto gli analisti, i manager delle banche d'investimento, fino a sopraffare anche l'allora presidente della Federal Reserve, Alan Greenspan. Però, sebbene siano stati commessi gravi errori di valutazione, che hanno determinato una differenza sostanziale tra i prezzi e i fondamentali delle Dot-com companies, è stata verificata l'impossibilità, durante quel periodo, di effettuare operazioni di arbitraggio prive di rischio, in quanto, solamente dopo lo scoppio della bolla, è stato possibile realizzare che i notevoli valori di mercato assegnati alle aziende del settore informatico erano incoerenti con una valutazione razionale e che le proiezioni di crescita previste erano insostenibili. Appurato, dunque, che, durante la bolla, non sono state disponibili opportunità sicure e redditizie di arbitraggio, e che, dopo il crollo, i prezzi delle azioni si sono, finalmente, adeguati a livelli che riflettevano più ragionevolmente il probabile valore attuale dei loro flussi di cassa futuri attesi,

bisogna però ammettere che i prezzi delle attività in questione sono rimasti ad un livello non consono per quasi un quinquennio. Pertanto, il mercato azionario potrebbe aver temporaneamente fallito nel suo ruolo che consiste nel garantire un'efficiente allocazione del capitale. Fortunatamente, i periodi in cui si verificano delle bolle speculative sono l'eccezione piuttosto che la regola.

1.6 CONSIDERAZIONI RIEPILOGATIVE

Molti attacchi alle implicazioni dell'EMH sembrano più che giustificati alla luce delle evidenze empiriche e delle anomalie osservate nei mercati finanziari, quali crisi, bolle speculative e fenomeni di gregge. Dunque, a causa dell'esistenza di patterns e imperfezioni, del fatto che il prezzo non rifletta sempre il valore fondamentale sottostante, e che le informazioni disponibili risultino distorte dall'irrazionalità di alcuni operatori che riflettendosi nei prezzi viene propagata nel mercato, sembra impossibile affermare che i mercati siano efficienti in maniera assoluta. Nonostante ciò, queste anomalie, anche se esistono, spesso risultano osservabili solamente dopo essersi manifestate, non permettendo così di ottenere rendimenti in eccesso aggiustati per il rischio costantemente superiori alla media. Possono esserci delle opportunità di arbitraggio, ma queste sono destinate ad essere non sistematiche, in quanto proprio le strategie di investimento volte a sfruttare le inefficienze sono in realtà l'espediente che permette di mantenere efficiente il mercato, rendendo gli errori di valutazione temporanei e riportando i prezzi verso un valore più in linea con i fondamentali. Infatti, come afferma l'economista Benjamin Graham nel suo libro del 1965 intitolato "The Intelligent Investor", il mercato azionario nel breve periodo può essere un meccanismo di voto, ma nel lungo periodo è un meccanismo di ponderazione. Un'ulteriore riflessione sul tema è stata fatta dall'economista e portfolio manager Richard Roll, il quale, durante il First Annual Symposium on Global Financial Markets, tenutosi ad Ann Arbor, in Michigan, tra il 3 e il 4 ottobre 1991, affermò che, pur avendo provato ad investire il suo denaro e quello dei suoi clienti, affidandosi ad una molteplice varietà di strategie basate sulla previsione di anomalie documentate da rinomate ricerche accademiche, non era ancora riuscito ad ottenere alcun profitto in maniera sistematica. Secondo Roll, se non vi è nessuna opportunità che gli investitori riescano a sfruttare in modo sistematico, allora non si può affermare che le informazioni non siano state correttamente incorporate nei prezzi delle azioni.

Negli ultimi anni, il significativo aumento della popolarità dei fondi indicizzati che seguono i principali indici di mercato, sia fondi comuni di investimento che ETF, è dovuto, almeno in parte, all'ampia accettazione popolare dell'ipotesi dei mercati efficienti. Infatti, gli investitori che sottoscrivono la EMH sono più inclini a investire in fondi indicizzati passivi, progettati per rispecchiare la performance complessiva del mercato, e meno propensi a pagare commissioni elevate per la gestione di fondi da parte di esperti, poiché ritengono che persino i migliori gestori di fondi possano garantire loro rendimenti superiori in modo significativo a quelli medi di mercato. Inoltre, con l'avvento di sistemi computerizzati sempre più potenti e veloci, che permettono investimenti automatizzati sulla base di rigorosi metodi analitici, matematici o fondamentali, e con l'interconnessione globalizzata, che garantisce una disponibilità immediata dell'informazione a tutti coloro che la richiedano, i prezzi si adattano sempre più rapidamente alle notizie rilevanti e, dunque, la rilevanza dell'ipotesi di efficienza dei mercati potrebbe aumentare ulteriormente.

2. CRIPTOVALUTE

Nel corso dell'ultimo decennio, le nuove tecnologie e l'evoluzione esponenziale della rete internet hanno determinato diverse innovazioni, che, favorite dai progressi della crittografia, ovvero dell'applicazione di metodi atti a rendere un messaggio comprensibile solamente a coloro che sono autorizzati a leggerlo, hanno apportato un cambiamento radicale nell'economia globale, con particolare riferimento al settore finanziario, sotto il profilo delle modalità di scambio di beni, servizi e di ogni attività finanziaria in generale.

Tra le più significative applicazioni della tecnologia digitale al settore finanziario spicca la nascita e la diffusione delle criptovalute, la più nota delle quali è Bitcoin. È prassi comune utilizzare la b maiuscola per riferirsi al network e la b minuscola per riferirsi alle unità di criptovaluta.

Però, ci sono stati diversi tentativi di creazione di moneta digitale prima dell'invenzione di Bitcoin, annunciato per la prima volta al pubblico il 31/10/2008. Perciò, è possibile affermare che la creazione di Bitcoin non è stata un evento unico, frutto della genialità di una singola persona, quanto, piuttosto, il risultato di un cammino che iniziò con la diffusione delle tecnologie digitali e che, in un processo continuo di prove ed errori, ha provato a giungere ad una soluzione in grado di bilanciare efficienza, sicurezza e decentralizzazione.

Infatti, già nel 1983, David Chaum, un informatico americano, pubblicò un saggio intitolato "Advances in Cryptology: Proceedings of Crypto 82", in cui, nella sezione "Blind Signatures for Untraceable Payments", cercò di formalizzare il sistema sottostante la crittografia e un possibile sistema di pagamento anonimo. Nel 1990 fondò la compagnia DigiCash e cinque anni più tardi diede vita alla prima valuta digitale denominata eCash, grazie alla quale introdusse una primitiva tecnica crittografica di "blind signature" per rendere i pagamenti anonimi e sicuri: per evitare che passassero delle informazioni indesiderate, egli utilizzò un sistema che definì "blinding formula", che da un lato confermava la legittimità e l'esistenza della valuta digitale, ma dall'altra ne impediva la tracciabilità. Non essendo, però, provvista di una tecnologia blockchain a supporto, eCash dipendeva da un'autorità centralizzata che ne garantiva l'emissione, ovvero la società stessa creata da David Chaum. Nel 1999, DigiCash fallì a causa di malumori interni all'azienda, dovuti alle continue declinazioni da parte di Chaum delle offerte di acquisto proposte da ING prima e da Microsoft in un secondo momento.

A seguito delle idee esposte da David Chaum, verso la fine degli anni '80, la crittografia e le tecnologie di miglioramento della privacy hanno cominciato ad essere viste come strumenti con un'importanza significativa al fine di suscitare cambiamenti nella realtà sociale e politica.

Infatti, nel 1992, Eric Hughes, Timothy May e John Gilmore, rispettivamente programmatore informatico, ingegnere elettronico e co-fondatore di "Cygnus Support", una piattaforma che aveva lo scopo di supportare a livello commerciale i software gratuiti, fondarono un movimento denominato Cypherpunks. Coloro che facevano parte di questo movimento ritenevano essenziale il tema della privacy, della libertà di parola e del diritto di ogni individuo di rivelare la propria identità solo quando desiderato, soprattutto a causa dello sviluppo sempre più prorompente dell'era digitale. Inoltre, i Cypherpunks credevano che le istituzioni governative non potessero garantire la privacy delle persone, in quanto tale necessità si discostava dal loro interesse. Per questo motivo, hanno sviluppato sistemi di transazione anonimi, per difendere la privacy tramite la crittografia.

Nel 1998, l'informatico e crittografo Nick Szabo cercò di lanciare BitGold, una valuta digitale decentralizzata, nel cui progetto, per la prima volta, si poneva l'accento sulla distribuzione

comunitaria della gestione della valuta e sulla creazione di un algoritmo proof-of-work per limitarne la produzione ed evitare fenomeni inflazionistici.
Nello stesso anno, l'ingegnere informatico Wei Dai, membro dei Cypherpunks, ideò un sistema di pagamento che doveva essere distribuito, digitale ed anonimo e che avveniva attraverso due canali di trasmissione, uno anonimo ed uno pubblico, la cui sincronicità verificava la transazione. Egli sviluppò due protocolli per una moneta digitale, chiamata b-money, le cui transazioni sarebbero state protette da algoritmi crittografici. In entrambi i protocolli, annunciava che la creazione di b-money sarebbe stata affidata alla proof-of-work, ovvero alla soluzione di un problema computazionale complesso, ancora insoluto, e che le transazioni, per le quali era necessaria la presenza di un mediatore, sarebbero state regolate da un contratto, in cui doveva essere indicato l'importo massimo dovuto in caso di inadempienza, ovvero nel caso in cui uno degli utenti coinvolti nella transazione non fosse stato in grado di adempiere al suo dovere di pagamento verso le altre parti. La differenza tra i due protocolli risiede in un'unica caratteristica, ovvero nel numero di partecipanti che devono tenere il conto di quanti b-money sono posseduti da ciascuna entità che fa parte del network: nel primo protocollo, che lo stesso Wei Dai definì impraticabile, il conto doveva essere custodito da ogni partecipante in database separati, mentre nel secondo protocollo, il conto doveva essere custodito solo da un sottoinsieme dei partecipanti, chiamati server.
Nel white paper di Bitcoin, ovvero nel documento informativo che espone in modo dettagliato le caratteristiche di Bitcoin, Satoshi Nakamoto, pseudonimo di una persona la cui identità è ancora sconosciuta, fece riferimento al testo in cui Wei Dai espone questi due protocolli, proprio perché essi rappresentano la base di partenza per lo sviluppo di Bitcoin.
Secondo Satoshi Nakamoto, il sistema di acquisti elettronici su internet presentava una serie di debolezze, in quanto fondato sulla fiducia che ogni partecipante doveva rispettivamente riporre nella controparte: da un lato i venditori dovevano esigere dai propri clienti varie informazioni per mantenere un livello di fiducia elevato, e dall'altro i compratori dovevano fidarsi dei venditori, non potendo utilizzare una moneta fisica per fare i loro acquisti. Molti problemi legati a questo tipo di transazioni possono, però, essere superati attraverso la creazione di un sistema di pagamenti basato su un moneta digitale crittografata, con software open source e rete peer-to-peer, senza la necessità di un intermediario tra le transazioni. Il peer-to-peer è un modello di architettura informativa in cui non esiste la tipica gerarchia server-client ma tutti i nodi sono paritari, per cui le funzioni di server e utente sono intercambiabili, ovvero ogni terminale è client e server al tempo stesso. La crittografia, ovvero la scienza che si occupa della cifratura e della protezione generale di dati e informazioni, risulta indispensabile per un sistema di pagamento senza contanti e interamente digitale, la cui funzionalità non dipende da un'istanza centrale e regolatrice. Nelle criptovalute vengono impiegate, principalmente, due procedure crittografiche: funzioni hash e firme digitali. Le funzioni hash sono necessarie a verificare l'integrità dei dati e a codificare gli indirizzi di conto e le operazioni dei partecipanti, costituendo la base della blockchain, mentre le firme digitali consentono di dimostrare lo stato delle informazioni cifrate senza esporle.
Nel white paper del 31/10/2008, pubblicato sulla mailing list metzdowd.com ed intitolato "Bitcoin: A Peer-to-Peer Electronic Cash System", Satoshi Nakamoto descrisse il suo progetto come un sistema di pagamento elettronico basato, invece che sulla fiducia, sulla prova crittografica, presentata sotto forma di transazioni verificate e convalidate in un registro aperto e distribuito chiamato blockchain. Inoltre, questo nuovo sistema di pagamento elettronico, come affermato il 09/01/2009, quando venne annunciata la prima versione di Bitcoin sulla piattaforma di sviluppo SourceForge, aveva come caratteristiche la decentralizzazione, ovvero l'assenza di un server o di

un'autorità centrale, e aveva come scopo quello di evitare il fenomeno del cosiddetto double-spending, ovvero la possibilità che una stessa transazione potesse essere utilizzata per due pagamenti diversi. Il primo blocco, noto come genesis block, da cui si è sviluppata in seguito la blockchain di Bitcoin, è stato realizzato il 03/01/2009, e il blocco successivo ha registrato la prima transazione di bitcoin tra Satoshi Nakamoto e l'attivista crittografo Hal Finney. Il 05/10/2009 è stato stabilito il primo tasso di cambio tra bitcoin e il dollaro americano, secondo cui un dollaro era equivalente a 1309 bitcoin.

Dal 2011, con la creazione di Litecoin, è iniziato un massiccio processo di creazione di criptovalute alternative a Bitcoin, denominate perciò altcoin, alcune delle quali sono nate da nuovi progetti con una propria blockchain e funzionalità specifiche, mentre altre sono il risultato di hardfork, ovvero di aggiornamenti della blockchain esistente, implementati da un determinato nodo della rete di una specifica criptovaluta. Quando tali aggiornamenti risultano incompatibili con il protocollo blockchain esistente, avviene una scissione permanente della criptovaluta iniziale in due diverse entità, ognuna delle quali propagherà le proprie transazioni e i propri blocchi, come è accaduto ad esempio per Bitcoin nel 2017 con la creazione di Bitcoin cash, o per Ethereum durante l'anno precedente con la creazione di Ethereum Classic.

Il mercato delle criptovalute, il cui significato va esteso al più generale concetto di crypto-asset, ovvero dei beni digitali registrati su un sistema distribuito e garantito dalla crittografia, considerando così, oltre alle vere e proprie criptovalute, anche i token e le stablecoin, si è sviluppato enormemente, dimostrando in dieci anni una crescita esponenziale, e la sua dimensione può essere riassunta nella capitalizzazione totale di mercato, che, come si può notare in figura 8, si attesta in data 10 febbraio 2021 intorno ai 873,5 miliardi di dollari.

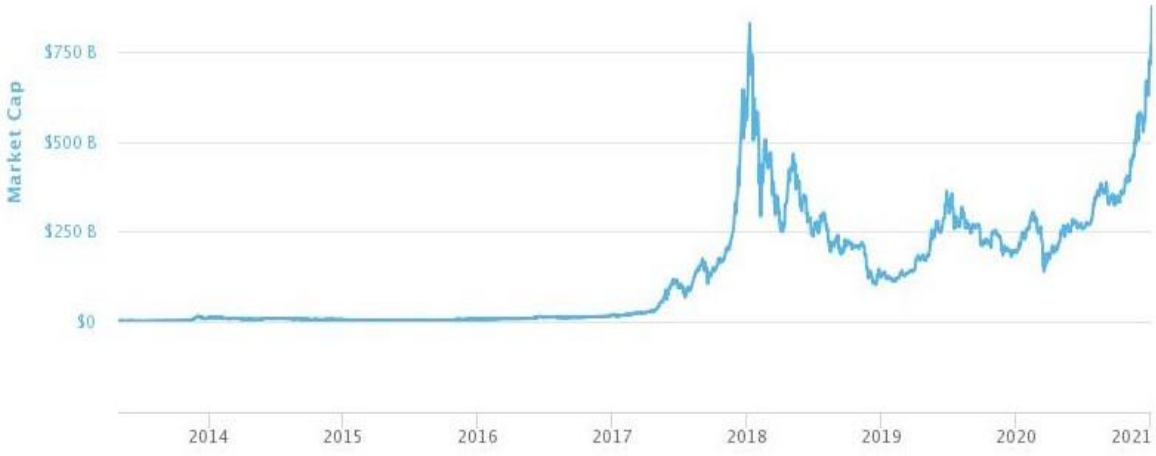

Figura 8: capitalizzazione totale del mercato delle criptovalute (fonte: coinmarketcap.com)

2.1 DEFINIZIONE DI CRIPTOVALUTA

Stabilire una definizione universale di criptovaluta risulta particolarmente arduo, a causa della novità, del cambiamento costante, della rapida evoluzione che caratterizza questo nuovo mercato, e della confusione che aleggia attorno a questo tema. Inoltre, le definizioni che vengono stabilite possono essere soggette ad aggiornamenti, a seguito della trasformazione continua dell'ecosistema

delle criptovalute. Dalla nascita di Bitcoin nel 2008, le maggiori istituzioni, a livello europeo ed internazionale, hanno effettuato analisi approfondite relative alle criptovalute, toccando l'argomento da diversi punti di vista e provando sempre a darne una definizione, basandosi sulle caratteristiche e sulle funzionalità di questa innovazione.
La Banca Centrale Europea (BCE), nel rapporto del 2012 dal titolo "Virtual Currency Schemes", ha classificato le criptovalute come un sottoinsieme delle valute virtuali, definendo queste ultime come una forma di moneta digitale non regolamentata, emessa e controllata solitamente da coloro che l'hanno sviluppata, e utilizzata ed accettata dai membri di una specifica comunità virtuale. Secondo lo studio effettuato, le valute virtuali possono essere distinte in tre tipologie, in base all'interazione con le valute tradizionali e con l'economia reale:

- Valute virtuali chiuse: questa tipologia di valute virtuali può essere utilizzata solamente all'interno di un sistema virtuale chiuso, come ad esempio nei giochi online, senza la possibilità di essere scambiata con le valute tradizionali.
- Valute virtuali unidirezionali: questa tipologia può essere acquistata tramite valute tradizionali, grazie all'esistenza di un tasso di conversione, e successivamente può essere utilizzata per acquistare beni e servizi digitali.
- Valute virtuali bidirezionali: per questa tipologia esiste un tasso di conversione sia per l'acquisto che per la vendita, rendendo tali valute bilateralmente collegate all'economia reale. La valuta virtuale può essere successivamente utilizzata per acquistare beni e servizi sia digitali che reali.

Le criptovalute, come il bitcoin, sono valute virtuali di quest'ultimo tipo, in quanto possono essere convertite in valute a corso legale e viceversa e possono essere utilizzate per acquistare beni e servizi sia digitali che reali.
Nel 2015, a causa degli sviluppi e della maggiore comprensione del tema in analisi, la BCE ha proposto un'ulteriore definizione di valute virtuali, aggiornando quella precedentemente data. Secondo il rapporto, intitolato "Virtual Currency Schemes – a further analysis", le valute virtuali sono rappresentazioni digitali del valore, non emesse da una banca centrale, da un istituto di credito o da un istituto di moneta elettronica, che in alcune circostanze possono essere utilizzate come alternativa alla moneta. All'interno di questo paper, le criptovalute sono state definite come una valuta virtuale bidirezionale decentralizzata.
Nel 2016, all'interno del report pubblicato dall' IMF Staff Team, dal titolo "Virtual Currencies and Beyond: Initial Considerations", anche il Fondo Monetario Internazionale (FMI) ha classificato le criptovalute come un sottoinsieme di valute digitali, come si può notare in figura 9.
Secondo il FMI, le valute digitali sono rappresentazioni digitali di valore, emesse da sviluppatori privati e denominate nella propria unità di conto, e tale concetto copre una più ampia gamma di valute, che vanno da contratti di debito informali, detti anche IOU, come ad esempio le miglia aeree grazie a cui i viaggiatori che le accumulano ricevono sconti da parte delle compagnie aeree, a valute virtuali garantite da asset esistenti e tangibili quali l'oro o da valute nazionali e dal merito di credito dell'emittente, per finire con le criptovalute come il bitcoin.
Il Committee on Payments and Market Infrastructures (CPMI), un organismo della Bank for International Settlements (BIS), che promuove, controlla e formula raccomandazioni sulla sicurezza e sull'efficienza dei pagamenti, delle liquidazioni, dei regolamenti e dei relativi accordi, sostenendo così la stabilità finanziaria e l'economia in generale, ha elaborato un report nel 2015 dal titolo "Digital Currencies". All'interno di tale studio, le valute digitali sono state definite come asset aventi alcune caratteristiche tipiche della moneta, come ad esempio il fatto di essere utilizzate come

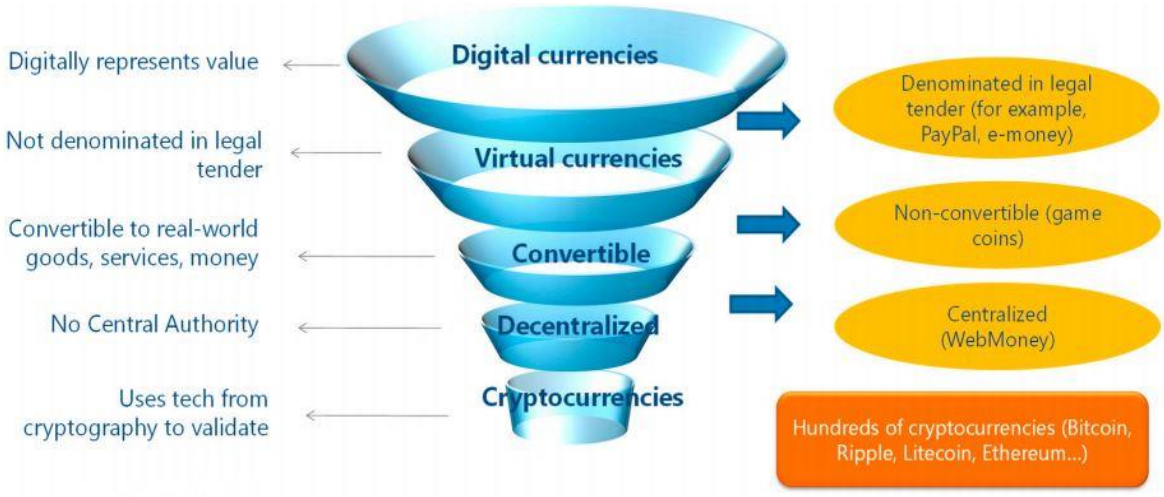

Figura 9: rappresentazione delle criptovalute come sottoinsieme delle valute virtuali

mezzo di pagamento, ma non emessi e non garantiti da istituzioni monetarie specifiche, con un valore intrinsecamente nullo, determinato solamente dalla domanda e dall'offerta e dalla fiducia nel poter scambiare tali asset con beni, servizi o con un determinato ammontare di valuta avente corso legale, e che si avvalgono di distributed ledgers per consentire scambi peer-to-peer a distanza in assenza di fiducia tra le parti e senza la necessità di intermediari. Il CPMI afferma che, nel linguaggio comune, ci si riferisce a tali asset con il termine criptovalute, per mettere in risalto l'utilizzo della crittografia, quale tecnica per la loro emissione e per la validazione delle transazioni.
Nel 2014, l'European Banking Authority (EBA), nell'elaborato dal titolo "EBA Opinion on 'virtual currencies'", si riferisce alle criptovalute come a valute virtuali, definite come rappresentazioni digitali di valore che non sono emesse da una banca centrale o da un'autorità pubblica né necessariamente collegate a una valuta fiat, ma che sono utilizzate da persone fisiche o giuridiche come mezzo di scambio e possono essere trasferite, conservate o scambiate digitalmente.
La European Securities and Markets Authority (ESMA) ha fatto riferimento alle criptovalute come valute virtuali, in un avviso paneuropeo emesso il 12/02/2018 in collaborazione con la European Insurance and Occupational Pensions Authority (EIOPA) e l'EBA. Coerentemente con quanto affermato dall'EBA, le valute virtuali sono definite come rappresentazioni digitali di valore che non sono né emesse né garantite da una banca centrale o da un'autorità pubblica e non hanno lo status legale di valuta o di denaro.
La Banca Mondiale, nel documento del 2017 intitolato "Distributed Ledger Technology (DLT) and blockchain", ha classificato le criptovalute come un sottoinsieme di monete digitali, che definisce come rappresentazioni digitali del valore che sono denominate nella propria unità di conto, distinte dalla moneta elettronica, che è semplicemente un meccanismo di pagamento digitale denominato in moneta fiat. A differenza della maggior parte delle altre istituzioni, la Banca Mondiale non ha trattato le criptovalute come un prodotto secondario e derivato da un sottoinsieme delle valute virtuali, ma si è soffermata su tale argomento, definendo le criptovalute come valute digitali che si basano su tecniche crittografiche al fine di raggiungere il consenso.
La Financial Action Task Force (FATF), un organismo intergovernativo indipendente che sviluppa e promuove politiche per la protezione del sistema finanziario globale, nel report del 2014 dal titolo "Virtual Currencies - Key Definitions and Potential AML/CFT Risks", ha definito le valute virtuali come rappresentazioni digitali di valore che possono essere scambiate digitalmente e che

funzionano come mezzo di scambio e/o unità di conto e/o deposito di valore, ma che non hanno corso legale, non sono emesse né garantite da nessuna giurisdizione, e che possono assolvere alle funzioni precedentemente elencate solamente a seguito di accordi tra coloro che aderiscono alla comunità che accetta tali valute virtuali. Inoltre, la FATF, riprendendo la classificazione operata dalla BCE nel 2012, suddivide le valute virtuali in convertibili e non convertibili, facendo rientrare le criptovalute nel primo raggruppamento. Secondo la FATF, dunque, le criptovalute sono valute virtuali convertibili, decentralizzate, basate sulla matematica, protette dalla crittografia, open-source, che non hanno un'autorità amministrativa centrale di monitoraggio e supervisione, e che permettono scambi e transazioni peer-to-peer.

La conclusione principale, che si può trarre dalle diverse prospettive sopra esposte, è che non esiste una definizione generalmente accettata del termine criptovalute. Sebbene tutte le istituzioni prese in considerazione ritengano le criptovalute come un sottoinsieme delle valute digitali o virtuali, la maggior parte di esse si è astenuta dal definire il termine nel suo complesso. Solamente la Banca Mondiale e la FATF hanno proposto una definizione chiara.

Cercando di riassumere le definizioni avanzate dagli organi istituzionali citati, si può affermare che una criptovaluta è una rappresentazione digitale del valore, destinata a costituire un'alternativa peer-to-peer alle monete a corso legale emesse dagli Stati, utilizzata come mezzo di scambio, indipendente da qualsiasi banca centrale, decentralizzata, garantita da un meccanismo noto come crittografia e che può essere convertita in valuta avente corso legale e viceversa.

Inoltre, il termine criptovalute, che viene spesso usato in senso molto ampio, va distinto da altri due concetti, token e cryptosecurities, che rientrano nella sfera dei crypto-asset, ovvero dei beni digitali registrati su un sistema distribuito e garantito dalla crittografia. I token sono crypto-asset che offrono una funzionalità diversa da quella del generico mezzo di scambio, in quanto, emessi nell'ambito di una Initial Token Offering (ITO) al fine di raccogliere fondi per un determinato progetto, rappresentano una sorta di credito nei confronti di un'entità, e possono, dunque, essere utilizzati per l'acquisto di determinati prodotti o servizi, ma solamente sulla piattaforma del token stesso. I token sono rappresentazioni digitali di beni o diritti reali, che, sebbene abbiano le stesse caratteristiche delle criptovalute a livello di sicurezza, non sono, però, "nativi" della blockchain sulla quale vengono memorizzate le transazioni che lo riguardano, in quanto utilizzano il registro delle transazioni di un'altra criptovaluta. Alcuni token possono essere paragonati a strumenti finanziari tradizionali, come azioni o obbligazioni, e sono comunemente chiamati "security token" o "investment token". Altri token, denominati "utility token", garantiscono ai loro detentori l'accesso futuro a specifici prodotti o servizi. I cryptosecurities, invece, rappresentano titoli societari che vengono registrati, emessi o traferiti tramite la tecnologia blockchain, assicurando il processo con la crittografia, affinché la tabella di capitalizzazione di una società sia sempre accurata ed aggiornata.

Un'ulteriore analisi compiuta dai grandi organi istituzionali riguarda la possibilità di considerare le criptovalute come vere e proprie valute, alla stregua delle valute aventi corso legale, e dunque sull'accuratezza stessa del termine criptovalute, sebbene questo sia ormai diventato di uso comune. Affinché le criptovalute vengano considerate come vere e proprie valute, ad esse vanno riconosciute le stesse funzionalità a cui assolvono le monete a corso legale, ovvero le funzioni di unità di conto, di mezzo di pagamento comunemente accettato e di deposito di valore.

Secondo le BCE, come affermato in un paper del 2019 dal titolo "Crypto-Assets: Implications for financial stability, monetary policy, and payments and market infrastructures", il termine criptovalute va sostituito con crypto-asset, in quanto l'assenza di un'istituzione specifica, come una

banca centrale o un'autorità monetaria, che protegga e regoli il valore dei crypto-asset ostacola il loro uso come forma di denaro, poiché la loro volatilità ne impedisce l'uso come riserva di valore, ne scoraggia l'uso come mezzo di pagamento e ne rende difficile l'uso come unità di conto. L'elevata volatilità delle criptovalute, causata dal fatto che il tasso di cambio in monete fiat sia basato solamente sulla domanda e sull'offerta di una determinata criptovaluta, produce fluttuazioni molto ampie nei prezzi, anche all'interno delle stesse giornate, rendendo, così, altamente inefficiente prezzare beni e servizi in unità di criptovalute, non consentendo, dunque, il corretto svolgimento della funzione di unità di conto. Per quanto riguarda la funzione di riserva di valore, l'elevata volatilità che caratterizza le criptovalute, non permette un adeguato assolvimento di tale funzione, ma è necessario effettuare una diversificazione geografica quando si considera questo argomento, poiché nei paesi, come ad esempio Venezuela e Turchia, in cui la moneta nazionale è soggetta ad una forte svalutazione causata dall'iperinflazione e dall'instabilità dell'economia, molti cittadini utilizzano le criptovalute, come ad esempio bitcoin, proprio per proteggersi dalla perdita di valore che subisce la moneta a corso legale: essi accettano la volatilità di bitcoin poiché ritengono che questa sia molto meno pericolosa e dannosa rispetto all'inflazione che colpisce i loro portafogli. Infine, se le unità di una criptovaluta possiedono un determinato controvalore, convertibile in valute fiat, il presupposto per l'uso come mezzo di pagamento è sostanzialmente soddisfatto, ma, per poter effettivamente pagare con le criptovalute, è necessario che anche il venditore le accetti come mezzo di pagamento. Perciò, si assiste ad un aumento della funzione di mezzo di pagamento man mano che sempre più persone entrano a far parte di questo mercato, accettando le criptovalute come mezzo di pagamento e facilitando il loro uso nella vita quotidiana, come sta avvenendo con PayPal, che dal 29/10/2020 permette ai propri clienti statunitensi di effettuare acquisti e vendite di alcune criptovalute direttamente tramite l'applicazione, e che, nel corso del 2021, ha come obiettivo quello di consentire a coloro che usano il servizio di usare le criptovalute proprio come mezzo di pagamento presso i negozi che accettano i pagamenti tramite PayPal, la quale si occuperà della conversione immediata della criptovaluta scelta per la transazione nella valuta a corso legale.

La Federal Reserve Bank di St. Louis, parte integrante della Banca Centrale degli Stati Uniti, in un articolo del 2018 intitolato "The Case for Central Bank Electronic Money and the Non-case for Central Bank Cryptocurrencies", considerando Bitcoin come esemplificativo del mercato delle criptovalute in generale, afferma l'unicità delle criptovalute, proponendo una differenziazione di questo nuovo asset rispetto alle esistenti forme di denaro, tramite l'utilizzo di una rappresentazione grafica denominata "control structure", riportata in figura 10. In quest'analisi vengono considerate tre dimensioni: la rappresentazione del denaro, che può essere fisica o virtuale, la gestione delle transazioni, che provoca una suddivisione tra sistemi di pagamento centralizzati e decentralizzati, e infine l'emissione del denaro, che può essere effettuata da un monopolio o può avvenire in condizioni di competizione e concorrenza.

Ad un vertice della control structure vi è il denaro contante (cash), rappresentato da un oggetto fisico, di solito una moneta o una banconota, il che significa che il suo valore è inseparabile dall'oggetto. Il detentore di un'unità di denaro contante è automaticamente il proprietario del valore corrispondente, e, di conseguenza, i diritti di proprietà delle unità di denaro contante, che circolano liberamente nell'economia, sono sempre chiaramente definiti senza che nessuno debba tenere traccia delle transazioni. Questa caratteristica consente un sistema di pagamento decentralizzato, in cui il contante può passare di mano tra due agenti senza l'intervento di terze

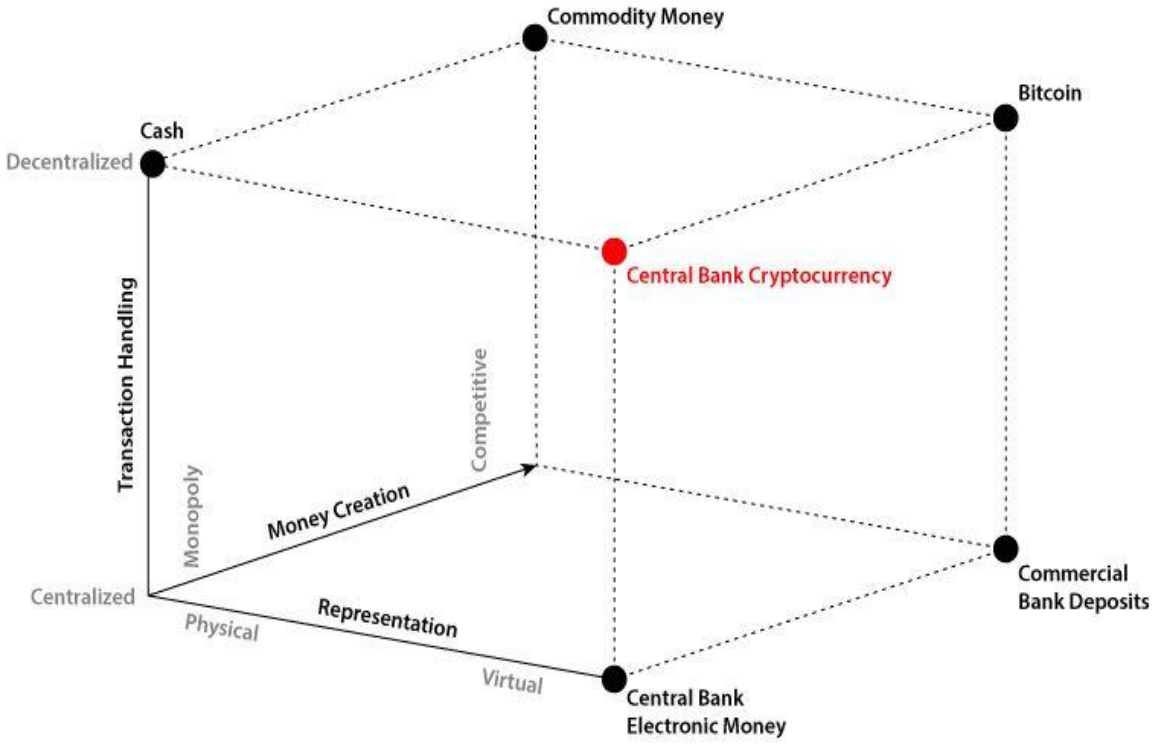

Figura 10: "control structure" dei diversi tipi di denaro

parti. Nella maggior parte dei paesi, la Banca Centrale o il Tesoro è l'emittente monopolistico del denaro contante.

Percorrendo l'asse dimensionale rappresentato dall'emissione della moneta, si trova poi la moneta merce (commodity money), ovvero quello strumento di pagamento rappresentato da un bene dotato di valore intrinseco proprio, come l'oro, che, ugualmente al denaro contante, è rappresentata da un oggetto fisico, e, dunque, anche al colui che detiene un'unità di moneta merce ha di default la proprietà dell'unità di valore, rendendo superflua la registrazione al fine di utilizzarla come strumento di pagamento. Il commodity money si differenzia dal contante per il suo processo di creazione competitivo, in quanto chiunque può entrare nel business dell'estrazione dell'oro, creandone nuove unità.

Vi sono, inoltre, i depositi presso le banche commerciali (commercial bank deposits), termine con cui si distinguono gli istituti di credito privati, che operano raccogliendo risorse finanziarie presso il pubblico ed erogando risorse finanziarie a titolo di credito, dalle banche d'affari. I depositi presso le banche commerciali, a differenza delle due tipologie di denaro esaminate in precedenza, sono denaro virtuale, e, a causa di questa loro natura, esistono solamente come registrazione in un sistema contabile. Quando viene effettuato un pagamento, i conti vengono aggiornati detraendo l'importo del pagamento dall'acquirente e accreditandolo al venditore. Nella maggior parte dei paesi, le famiglie e le imprese utilizzano i tali depositi per effettuare pagamenti elettronici, tramite carte di credito, carte di debito, assegni e online banking. Le banche commerciali competono tra loro nel formulare l'offerta più vantaggiosa al fine di attirare i depositi nella propria banca. Inoltre esse sono responsabili della gestione dei registri, perciò i depositi delle banche commerciali sono considerati come un sistema di pagamento centralizzato.

Tornando indietro sull'asse relativo all'emissione del denaro, si trova la moneta elettronica delle banche centrali (Central Bank electronic money), anch'essa virtuale, ma centralizzata ed emessa in

modo monopolistico. Nella maggior parte dei paesi, l'accesso pubblico alla moneta elettronica della Banca Centrale è limitato.
Nell'articolo in questione viene anche trattato il caso ipotetico di una criptovaluta emessa dalle banche centrali (Central Bank Cryptocurrency), ma la conclusione a cui giungono gli autori è che questa soluzione non avrebbe molto senso, in primo luogo poiché nessuna banca centrale rispettabile sarebbe incentivata ad emettere una moneta virtuale anonima che potrebbe essere usata per scopi illeciti e illegali, comportando un rischio reputazionale elevato per la banca stessa, e in secondo luogo poiché sorgerebbero problematiche legate alle normative KYC (know your customer) e AML (anti-money laundering), a cui le banche commerciali dovrebbero continuare a sottostare, ma che non verrebbero più rispettate dalle banche centrali. Inoltre, una criptovaluta emessa da una banca centrale perderebbe, per definizione, la caratteristica della decentralizzazione, proprio quella che rende le criptovalute così innovative.
Nell'ultimo vertice rimasto, vi sono le criptovalute, tra le quali bitcoin viene preso come punto di riferimento. Questa criptovaluta è il primo denaro virtuale, per il quale i diritti di proprietà delle varie unità monetarie sono gestiti in un sistema contabile decentralizzato definito blockchain, e la cui creazione di nuove unità risulta competitiva, poiché chiunque ha la possibilità di contribuire alla creazione di nuove unità di criptovaluta, scaricando il software rispettivo. La peculiarità, sottolineata in questo articolo, che rende uniche le criptovalute, è la capacità di combinare i vantaggi transazionali del denaro virtuale con l'indipendenza sistemica della gestione decentralizzata delle transazioni.
In conclusione, sebbene vi siano alcune divergenze nella definizione di criptovaluta e una naturale confusione su alcuni termini e sull'accezione delle criptovalute quali denaro vero e proprio, vista la giovinezza e la continua evoluzione di questo mercato, non vi è alcun dubbio sulla portata rivoluzionaria di questa innovazione, che ha il potenziale di perturbare l'attuale sistema finanziario e le infrastrutture di pagamento, interessando tutte le imprese e le istituzioni governative che si occupano della gestione delle registrazioni contabili.

2.2 CARATTERISTICHE DELLE CRIPTOVALUTE E BLOCKCHAIN

Per comprendere meglio la portata rivoluzionaria e il funzionamento di questa innovazione finanziaria, è necessario analizzarne le caratteristiche peculiari. Le criptovalute sono caratterizzate dai seguenti elementi costitutivi:

- Protocollo: è un insieme di regole sotto forma di codice informatico, che specifica il modo in cui possono essere effettuate le transazioni.
- Blockchain: è una sorta di libro mastro, ovvero un registro che conserva la storia delle transazioni in maniera immodificabile.
- Decentralized Network: è una rete decentralizzata di partecipanti, che, sul proprio server, aggiornano, conservano e consultano il Distributed Ledger delle transazioni, secondo le regole imposte dal protocollo. Ogni partecipante alla rete viene definito nodo.

La peculiarità, globalmente riconosciuta come una "disruptive innovation", delle criptovalute risiede nella blockchain. Per comprendere il concetto di blockchain nel suo complesso, è necessario, innanzitutto, analizzare le Distributed Ledger Technologies.
Le Distributed Ledger Technologies (DLT) possono essere definite come un insieme di sistemi caratterizzati dalla presenza di un registro distribuito, detto anche libro mastro distribuito o

condiviso, governato, tramite determinate modalità di gestione del consenso e di impostazione del registro stesso, in modo da consentire l'accesso e la possibilità di effettuare modifiche da parte di più nodi di una rete. Un libro mastro distribuito è un database di dati, che viene consensualmente condiviso su una rete peer-to-peer, senza la necessità di un'autorità centrale. Ogni partecipante diventa un testimone pubblico dei dati registrati sul libro mastro distribuito. Inoltre, tutti i partecipanti possiedono una copia identica dei dati sulla rete condivisa, e ciò rende quasi impossibile per una singola entità apportare modifiche al database esistente.
Le DLT rappresentano un'evoluzione dei Centralized e Decentralized Ledger, le cui differenze sono mostrate graficamente in figura 11. Gli archivi cartacei centralizzati sono diventati digitali e, successivamente, vi è stata un'accelerazione innovativa, grazie alla contemporanea disponibilità di due fattori abilitanti, quali la crittografia e lo sviluppo di algoritmi di controllo e verifica dei dati, che hanno permesso l'attuazione di una reale e completa logica distribuita, in cui non esiste più nessun centro e dove il concetto di governance è costruito attorno ad una nuova espressione di fiducia tra tutti i soggetti che partecipano alla rete. Nessuno ha la possibilità di prevalere e il processo decisionale passa rigorosamente attraverso un processo di costruzione del consenso.

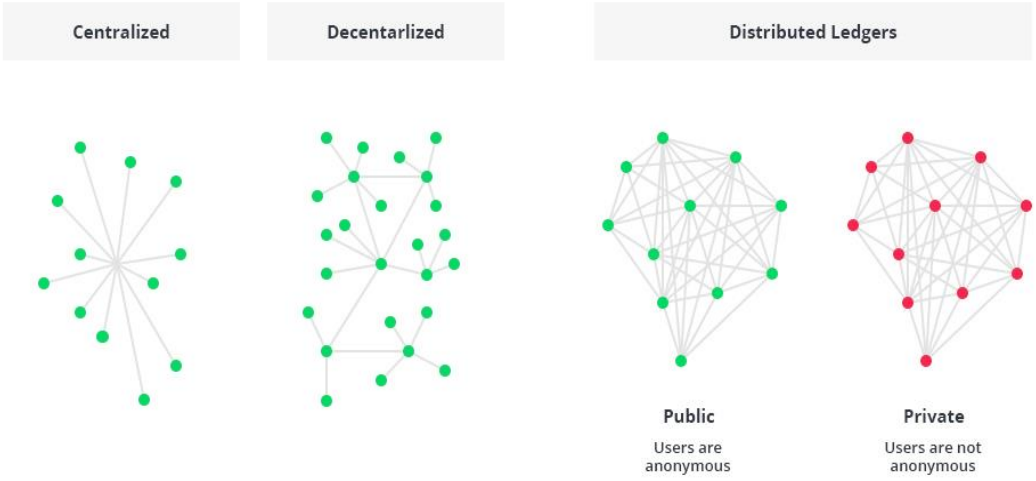

Figura 11: evoluzione dei Ledgers

Con le DLT, gli aggiornamenti del database delle transazioni non sono più gestiti, controllati e coordinati, come accade tradizionalmente, a livello centrale, ma sono invece creati e caricati da ciascun nodo in modo indipendente. Il presupposto fondamentale per il funzionamento di una DLT risiede nella creazione di grandi network, costituiti da una serie di partecipanti, ognuno dei quali deve gestire un nodo di questa rete. Ogni partecipante è, dunque, in grado di processare e controllare tutte le transazioni che avvengono nella rete, ma, allo stesso tempo, ogni singola transazione deve essere verificata e approvata dalla maggioranza dei partecipanti alla rete, in quanto l'autonomia, con cui ad ogni partecipante è permesso operare, è subordinata al controllo consensuale da parte degli altri nodi, ovvero al raggiungimento di un consenso sulle operazioni che vengono svolte, e, solamente a seguito di tale meccanismo di validazione, le operazioni sono poi autorizzate ed attivate.
Le blockchain è una tipologia di DLT, ma è solamente una delle diverse possibilità di gestione del consenso utilizzate per applicare le DLT: mentre tutte le blockchain sono dei libri mastro distribuiti, non tutti i libri mastro distribuiti sono blockchain. Inoltre, la blockchain non si limita ad essere utilizzata solo per le criptovalute, in quanto tale tecnologia può essere impiegata in vari settori e

può dunque avere una vasta gamma di applicazioni. Vi sono moltissimi ambiti applicativi, oltre alla finanza, in cui la blockchain è presente o lo sarà in futuro, come ad esempio le assicurazioni, la sanità, la pubblica amministrazione, la supply chain e altri ancora.
Per quanto riguarda le criptovalute, la tecnologia che ne garantisce lo sviluppo e il funzionamento è la blockchain definita permissionless o pubblica, in cui il registro, aperto e distribuito, è impostato e strutturato in modo tale da gestire le transazioni all'interno di una catena di blocchi, rappresentata in figura 12.

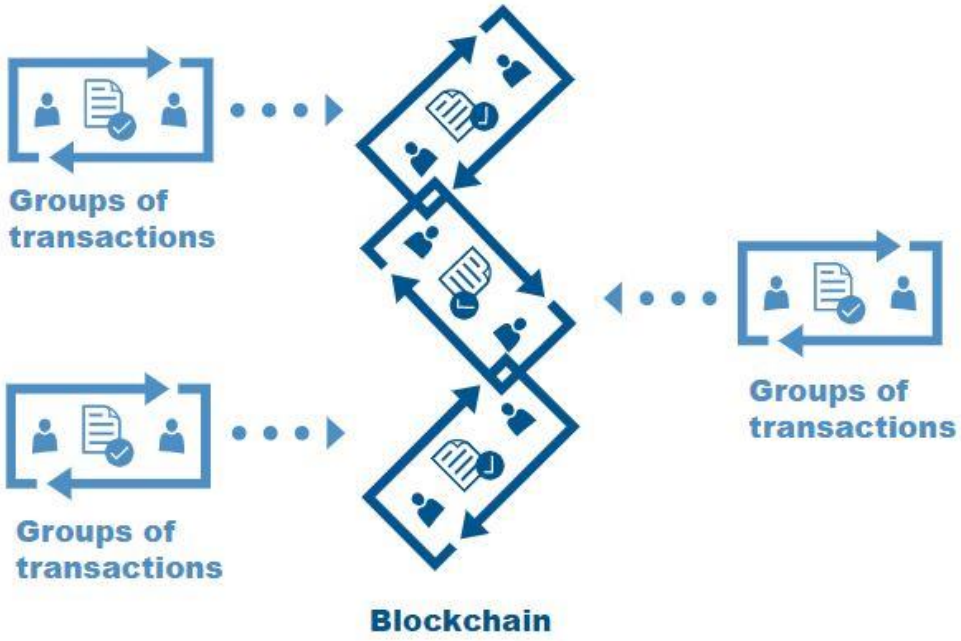

Figura 12: blockchain

La blockchain, di cui la maggior parte delle criptovalute possiede la propria, crea uno specifico tipo di libro mastro distribuito, che stabilisce un database immutabile condiviso da una rete decentralizzata, utilizzando la crittografia per convalidare e registrare tutte le transazioni attraverso un meccanismo di consenso, al fine di garantire l'integrità e la sicurezza dei dati. Dal punto di vista delle regole di gestione, ciascun blocco si aggiunge alla catena sulla base di un processo basato sul consenso distribuito su tutti i nodi della rete, che vengono chiamati a contribuire alla validazione delle transazioni presenti in ciascun blocco e alla loro inclusione nel registro. L'amministrazione del database spetta, dunque, ai partecipanti della rete peer-to-peer di base, che seguono un protocollo definito per convalidare le nuove operazioni. Di conseguenza, tutti i nodi scaricano automaticamente una copia integrale della blockchain, in modo tale che i dati riguardanti tutte le transazioni avvenute siano condivisi, accessibili e distribuiti tra tutti i partecipanti, e ciò permette una gestione dei dati in termini di verifica e di autorizzazione senza che sia necessaria un'istanza centrale per la supervisione delle operazioni avvenute. Ogni blocco di transazioni è collegato crittograficamente al suo predecessore per creare una catena di record di dati ininterrotti con indicazione del momento in cui sono stati creati, in modo tale che le transazioni siano registrate in modo ordinato e sequenziale.
Le blockchain su cui si basano le criptovalute sono definite permissionless ledger, in quanto sono archivi aperti, concepiti per non essere soggetti ad un controllo centralizzato, ovvero senza qualcuno che ne sia l'unico proprietario. Questo modello di blockchain, che permettere a ciascuno di contribuire all'aggiornamento dei dati e di disporre, in qualità di partecipante, della copia

immutabile di tutte le transazioni avvenute e approvate grazie al consenso, impedisce ogni forma di censura, in modo tale che nessuno abbia il potere di ostacolare una transazione o di invalidarla una volta che questa abbia conquistato il consenso necessario tra tutti i nodi della blockchain.
In figura 13 si può vedere il processo che una transazione di criptovalute deve seguire affinché venga completata.

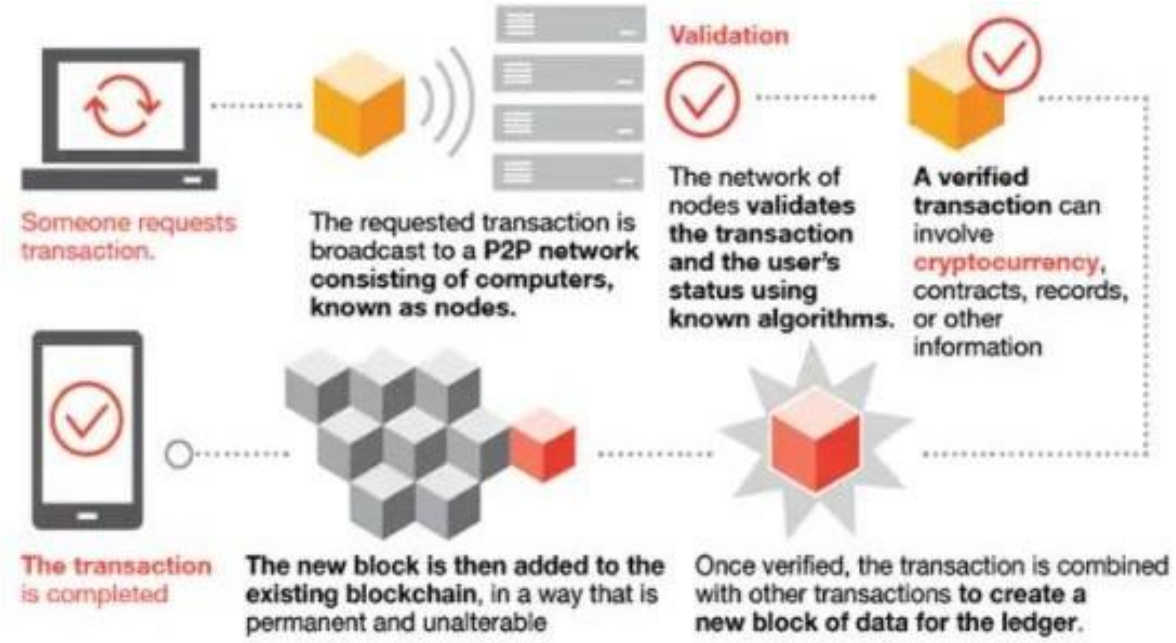

Figura 13: processo di una transazione tramite la blockchain

Una transazione è un semplice trasferimento di determinate unità di una criptovaluta da un partecipante del network ad un altro e contiene le informazioni relative all'indirizzo pubblico del ricevente, le caratteristiche della transazione, ovvero i valori oggetto di scambio e che necessitano di essere verificati, approvati ed archiviati, e la firma crittografica che garantisce sicurezza ed autenticità alla transazione. Qualunque transazione è sottoposta ad un meccanismo di firma a doppia chiave asimmetrica che funziona con un procedimento simile a quello della firma digitale, pur non essendo dotata di certificati rilasciati da organismi certificatori accreditati. La blockchain prevede l'impiego di algoritmi crittografici, che abilitano l'utente all'utilizzo del sistema tramite una chiave pubblica ed una privata. La chiave pubblica è visibile a tutti e può essere condivisa con chiunque, in quanto è utilizzata per ricavare gli indirizzi pubblici che identificano il portafoglio di un determinato partecipante su cui trasferire le unità di criptovaluta desiderate. Sebbene sia collegata al portafoglio digitale di un determinato partecipante, la chiave pubblica non permette di risalire all'identità della persona nel mondo reale, in quanto la chiave pubblica consiste in una semplice stringa di caratteri casuali. La chiave privata, invece, deve essere tenuta al sicuro, in quanto viene utilizzata per autenticare la proprietà delle criptovalute, per criptare il portafoglio digitale in cui queste ultime sono contenute e per sottoscrivere le transazioni. Questo framework permette all'utente di criptare le informazioni relative alla transazione, utilizzando la propria chiave privata, e di trasmetterle all'intero network, garantendo, al contempo, che solo il destinatario previsto, la cui chiave pubblica è stata inserita come indirizzo di destinazione, possa decifrarle, utilizzando la corrispondente chiave privata.

Dopo che la transazione è stata firmata, essa viene tramessa a tutti i nodi della rete, tramite la tecnologia peer-to-peer, ma, prima di essere confermata, la transazione deve passare attraverso un processo di verifica che rappresenta l'elemento fondamentale della blockchain.
Il processo di validazione della transazione, che viene unita ad altre transazioni per formare un blocco, prevede una fase di verifica e di approvazione basata su risorse che vengono messe a disposizione dai partecipanti alla blockchain e che sono finalizzate alla risoluzione di problemi matematici complessi, detti anche puzzle crittografici, permettendo, così, di disporre di un consenso distribuito e non più di un consenso basato su un intermediario o su un ente o istituzione centralizzata. I meccanismi di consenso, che si attivano una volta che la transazione necessita di verifica, sono definiti "proof of work" e "proof of stake". Questi meccanismi di consenso sono necessari a confermare che la transazione sia legittima e che non esista già nel sistema, garantendo che nessun pagamento venga addebitato più di una volta, evitando così il problema del double-spending.
La proof of work è un sistema che è diventato noto grazie all'associazione che viene fatta con Bitcoin e, perciò, spesso viene definito Consenso di Nakamoto. Nella proof of work, il consenso è rappresentato dalla regola definita "longest chain wins", secondo cui i partecipanti alla rete blockchain devono accettare la catena di blocchi più lunga come l'unica valida, impedendo, così, che catene multiple, ognuna delle quali rappresenterebbe una versione diversa della registrazione della storia delle transazioni, possano affiancarsi l'una all'altra. Affinché tale regola funzioni in modo sicuro, l'aggiunta di nuovi blocchi alla blockchain è progettata per far sì che la risoluzione del puzzle crittografico risulti sempre più dispendioso in termini di costi e di tempo. Dunque, man mano che la versione consensuale della blockchain diventa sempre più lunga, aumentano la potenza di calcolo e le risorse computazionali necessarie a convalidare nuovi blocchi. Coloro che partecipano alla risoluzione del problema crittografico e che, dunque, concorrono alla validazione del blocco e delle transazioni in esso contenute sono chiamati "miner", e il loro intervento, denominato "mining", che necessita di importanti risorse per essere svolto, sia in termini di potenza che di capacità elaborativa, viene incentivato e remunerato attraverso l'emissione di una unità di criptovaluta che dipende dalla tipologia di blockchain sulla quale viene eseguita la transazione. Il puzzle crittografico è concepito per mettere in competizione tutti i nodi, i quali contribuiscono alla risoluzione mettendo a disposizione la propria potenza di calcolo. Il nodo che riuscirà a risolvere il puzzle crittografico avrà il diritto di validare il blocco con la presentazione della proof of work, ovvero della prova che certifichi la risoluzione del puzzle, e otterrà la remunerazione. Tuttavia, tale protocollo di consenso può portare a significativi tempi di transazione, in quanto richiede una sofisticata potenza di calcolo per raggiungere il consenso. Infatti, Satoshi Nakomoto, nel white paper di Bitcoin, ha spiegato che la regola della longest chain non serve solo come prova della sequenza di transazioni che sono avvenute, ma anche come prova che tale sequenza provenga dal più grande pool di potenza computazionale. Da questo principio sono scaturite delle critiche che hanno portato i sostenitori dell'inclusione finanziaria e del decentramento a sostenere che, con la crescita della rete blockchain di Bitcoin, l'attività di mining sia diventata il privilegio quasi esclusivo di chi ha i mezzi per investire in hardware sempre più sofisticati e potenti, rimanendo competitivo. Inoltre, il mining, se effettuato da attori malintenzionati, sprovvisti di un interesse reale nel progetto di una determinata criptovaluta, e che detengono hardware, la cui potenza di capacità elaborativa e le cui risorse computazionali sono tali da risultare maggiori della metà dell'hash rate totale, dove l'hash rate rappresenta una misura della potenza di elaborazione di una rete blockchain, potrebbe, in tal caso, perturbare l'integrità e la sicurezza della blockchain. Infatti, tali attori malintenzionati

potrebbero perpetrare il cosiddetto “51% attack”, grazie al quale, controllando oltre la metà dell’hash rate totale e formando, così, la maggioranza nel meccanismo di consenso, potrebbero modificare l’ordine delle transazioni, impedire che alcune di esse siano confermate e persino fare in modo che si verifichi il fenomeno del double-spending. Questi attacchi sono più concretizzabili nelle blockchain in cui l’hash rate risulta concentrato tra pochi miners, o in cui è ridotto, in quanto, in tal caso, è più facile per un partecipante procurarsi la potenza di calcolo necessaria a controllare la maggioranza. Un 51% attack è avvenuto, ad esempio, nel gennaio 2019 sulla blockchain di Ethereum Classic.

Tali riflessioni, a cui si sono aggiunte considerazioni relative all’ingente consumo di energia elettrica in relazione all’epoca in cui viviamo, caratterizzata da un preoccupante cambiamento climatico, hanno generato profonde modifiche nei meccanismi di consenso del mondo delle criptovalute, a tal punto che alcune blockchain hanno abbandonato il meccanismo della proof of work per adottare, invece, quello della proof of stake. Questo è ciò che è successo per la criptovaluta Ether, la cui blockchain ha subito un aggiornamento, denominato Ethereum 2.0 (o anche Eth2 o Serenity), che è stato sperimentato per cinque anni e il cui lancio è avvenuto l’1/12/2020. Tale aggiornamento ha lo scopo di rendere maggiormente performanti la velocità, l’efficienza e la scalabilità del network, in modo che quest’ultimo possa elaborare un maggior numero di transazioni, evitando allo stesso tempo situazioni di sovraccarico. L’aggiornamento si articolerà in varie fasi scandite nel tempo e la prima che è stata attuata dalla data del lancio consiste proprio il passaggio dalla proof of work alla proof of stake. All’interno delle reti che adottano tale meccanismo, la sicurezza, il consenso e il mantenimento della rete stessa sono ottenuti tramite un processo definito staking, detto anche forging o minting, secondo cui i partecipanti devono dimostrare di essere in possesso delle criptovalute native di quella determinata rete, impegnandone una quota, che rimane inutilizzabile per un tempo asseganto, nel progetto della blockchain, al fine di competere per essere sorteggiati e avere il diritto a verificare una transazione. Dunque, il sistema di proof-of-stake funziona come una prova crittografica della proprietà di una determinata criptovaluta e come una prova di interesse per il successo del progetto. Un algoritmo deterministico seleziona in modo casuale, tra coloro che hanno messo in staking delle unità di criptovaluta, quali sono quei nodi che, in base alle quantità di criptovaluta detenuta e da quanto tempo la detengono, dovranno, di volta in volta, validare il nuovo blocco, e tali nodi saranno definiti “validator”. Altri validators possono poi attestare di aver verificato il blocco in questione e, quando ci sono abbastanza attestazioni, è possibile aggiungere il blocco alla blockchain. Questo processo di selezione randomizzata ha lo scopo di disincentivare i partecipanti dal tentativo di sabotare la storia delle transazioni e di minare il sistema della blockchain. Esattamente come avviene nella proof of work, coloro che convalidano un blocco vengono remunerati per il loro impegno, ma il meccanismo basato sulla proof of stake risulta più sostenibile dal punto di vista ambientale, in quanto non richiede l’eccessiva quantità di energia necessaria all’attività di mining. Inoltre, gli sviluppatori della proof of stake sostengono anche che il sistema è più resistente ai monopoli e alla centralizzazione del potere all'interno della rete, in quanto la partecipazione è disaccoppiata dal controllo sull'hardware e sulle risorse.

La validazione del blocco, sia per quanto riguarda la proof of work che per la proof of stake, richiede la risoluzione di un puzzle crittografico, che corrisponde all’applicazione al blocco da validare di un hash, ovvero di una funzione algoritmica informatica non invertibile che mappa una stringa di lunghezza arbitraria in una stringa di lunghezza predefinita, che, registrando tutte le informazioni relative al blocco attuale e incorporando l’hash con le informazioni relative al blocco precedente, identifica il blocco in maniera univoca e permette di creare la catena, collegando un blocco all’altro.

Tramite le funzioni hash, ciascun blocco diventa un archivio per tutte le transazione e per lo storico di ciascuna transazione, in modo tale che risulta possibile ripercorrere la catena di blocchi all'indietro.
Oltre al meccanismo di consenso, vi è un'altra operazione attraverso cui ogni blocco deve passare al fine di essere aggiunto alla blockchain, che consiste nell'applicazione da una marca temporale (timestamp) ad ogni transazione presente nel blocco. Il timestamp consente di associare una data ed un orario certi e legalmente validi alla transazione presa in considerazione, fissandone ed accertandone l'effettivo avvenimento, e definendo dunque una validazione temporale che può essere opponibile a terzi. La marca temporale è rappresentata in un formato che ne rende possibile la comparazione con altre date, permettendo di stabilire e definire un ordine temporale. Perciò, la marca temporale impedisce il fenomeno del double-spending e impedisce che la transazione, se correttamente eseguita, venga alterata o annullata.
Nel caso in cui il processo di verifica dovesse rilevare un errore, una anomalia, una discrepanza, il blocco viene rifiutato e tutti possono osservare il fatto che la transazione non è stata autorizzata. Se, invece, le informazioni relative alla funzione hash e al timestamp sono corrette, allora le transazioni che fanno parte di quel blocco vengono autorizzate, validate ed effettuate. Il nuovo blocco, che rappresenta ora il riferimento permanente, immutabile e immodificabile di quella specifica transazione viene, dunque, aggiunto alla catena di blocchi che forma la blockchain, e diventa accessibile a tutti i partecipanti, che possono consultarlo all'interno del proprio archivio, il quale viene aggiornato con la nuova transazione.
Finché una transazione non è confermata, è in sospeso e può essere falsificata, ma quando una transazione viene validata, diventa indelebile ed irreversibile, in quanto le stesse informazioni relative alle transazioni sono presenti su tutti i nodi e pertanto non possono essere retroattivamente alterate senza che vengano modificati tutti i blocchi successivi, se non attraverso un'operazione che richiede l'approvazione della maggioranza della rete.
La presenza di un registro aperto e condiviso tra tutti i nodi della rete è una delle caratteristiche principali della blockchain, che differenzia quest'ultima dagli archivi centralizzati, in termini di sicurezza. Infatti, se per modificare, danneggiare o distruggere un Central Ledger è necessario violare l'autorità centrale che lo gestisce, nel caso della blockchain bisognerebbe violare tutte le copie del database possedute da ogni partecipante alla rete e occorrerebbe farlo simultaneamente. Inoltre, non può esistere una blockchain falsa, in quanto tutti i partecipanti sono in possesso di un'unica versione autentica, che rappresenta lo storico delle transazioni e che può essere esibita per una verifica. Dunque, la blockchain risulta essere affidabile e sicura, in quanto non è gestita da un ente centrale ma tutti i nodi partecipano al controllo dell'intera catena con l'aiuto della crittografia, trasparente, poiché le transazioni sono visibili a tutti i partecipanti, conveniente in termini di costo, dal momento che si evitano le commissioni applicate dagli istituti finanziari tradizionali, necessari in tutte le transazioni convenzionali, e solida, poiché le transazioni non si possono alterare.
L'utilizzo di criptovalute basate sulla tecnologia blockchain permette perciò una maggiore velocità ed efficienza nei pagamenti, offre la possibilità di operare in qualunque momento senza l'influsso intermediatore delle istituzioni finanziarie, promuove l'inclusione finanziaria a livello globale permettendo anche a coloro che non hanno accesso all'attuale sistema bancario di partecipare all'economia globale, ed infine garantisce l'anonimato per coloro che effettuano transazioni, caratteristica che, però, ha reso le criptovalute molto attraenti per i criminali, che potrebbero utilizzarle per riciclaggio di denaro sporco e altre attività illegali.

2.3 STRUTTURA DEL MERCATO DELLE CRIPTOVALUTE

Il mercato delle criptovalute è composto da diversi attori e strumenti, di seguito elencati, che permettono al settore di funzionare e di svilupparsi:

- Utenti: sono coloro che, una volta ottenuta una determinata criptovaluta, la utilizzano per acquistare beni o servizi reali o virtuali da coloro che accettano pagamenti in criptovalute, oppure per effettuare pagamenti peer-to-peer, o infine per tenerla in portafoglio a scopo di investimento, ovvero in modo speculativo. Gli utenti possono ottenere le criptovalute acquistandole tramite una piattaforma di exchange con un pagamento in moneta fiat o in altra criptovaluta, oppure acquistandole attraverso una piattaforma di trading con uno scambio definito "P2P exchange", oppure facendo mining, rientrando così nella categoria dei miners, se la criptovaluta presa in considerazione è basata su un meccanismo di consenso proof of work, oppure accettando unità di criptovaluta come pagamento dei beni o servizi offerti, o come regalo o donazione da parte di altri utenti, e infine prendendo parte ad un'offerta iniziale gratuita o ad una vendita di massa organizzata dall'offerente della criptovaluta.
- Miners: come già esposto precedentemente, sono coloro che partecipano alla convalida delle transazioni sulle blockchain di criptovalute che si basano su un meccanismo di consenso proof of work, risolvendo un puzzle crittografico. I miners, che talvolta si raggruppano in "mining pool" al fine di raggruppare la potenza computazionale dei propri hardware al fine di renderla più performante, si occupano di garantire l'autenticità e la regolarità di ogni transazione e il primo tra loro che riesce a convalidare il blocco in questione riceve come ricompensa delle unità di criptovaluta appena coniate, attraverso una nuova emissione automatica e decentralizzata.
- Piattaforme di exchange: sono piattaforme online che offrono un servizio agli utenti, permettendo loro di scambiare monete fiat in criptovalute e viceversa o unità di una determinata criptovaluta in un'altra criptovaluta, solitamente dietro pagamento di una certa commissione. Vi sono exchange che permettono solamente la conversione tra diverse criptovalute, mentre altri consentono entrambe le modalità di scambio. Alcuni exchange sono dunque puri, garantendo solo il servizio per cui sono nati, mentre altri fungono anche da piattaforme di trading e spesso anche da custodial wallet. Esempi di piattaforme di exchange sono Binance, Coinbase, Bitfinex, Kraken, Bitpanda, Flamingo, Conio e molti altri. Le piattaforme di exchange su cui si acquistano e si vendono criptovalute non sono attualmente regolamentate, perciò non è prevista una tutela legale specifica per gli utenti in caso di contenzioso o fallimento della piattaforma. In un contesto di assenza di obblighi informativi e di regole di trasparenza, le piattaforme di exchange sono, inoltre, esposte a elevati rischi operativi e di sicurezza: esse, infatti, a differenza degli intermediari autorizzati, non sono tenute ad alcuna garanzia di qualità del servizio, né devono rispettare requisiti patrimoniali o procedure di controllo interno e gestione dei rischi, con conseguente elevata probabilità di frodi ed esposizione al cyber-crime.
- Piattaforme di trading: vi sono i P2P exchanges, detti anche exchanges decentralizzati, come ad esempio LocalBitcoins, le cui piattaforme non sono gestite da un'entità o da una società che supervisiona ed elabora tutte le operazioni, ma esclusivamente da un software, che si limita a collegare un acquirente con un venditore, permettendogli di condurre un'operazione di compravendita di criptovalute online, e poi vi sono le piattaforme di

trading più comunemente conosciute per le attività puramente speculative, come ad esempio eToro, che nascono come piattaforme per il trading su azioni e che stanno espandendo i loro servizi anche al mercato delle criptovalute.

- Wallet: sono portafogli digitali, per detenere, conservare e trasferire criptovalute. Più nello specifico, sono software che conservano le chiavi, pubblica e privata, le quali permettono di gestire le criptovalute possedute e di interagire con la rete blockchain. Un fornitore di wallet traduce la cronologia delle transazioni di criptovalute effettuate da un utente in un formato facilmente leggibile, che assomiglia molto a un normale conto bancario. Il wallet viene usato per generare un indirizzo pubblico corrispondente al wallet stesso, che serve per identificare pubblicamente colui che effettua o che riceve la transazione. L'indirizzo è una versione ridotta della chiave pubblica, in modo da renderne più semplice la condivisione. È possibile distinguere i wallet in due categorie: hot wallet e cold wallet. I primi, che a loro volta si suddividono in desktop, mobile e hybrid wallet, sono connessi a Internet e possono essere custodial o non-custodial. Per quanto rigurda i custodial wallet, è il fornitore del wallet che detiene la chiave privata per conto del proprietario del wallet, mentre, al contrario, per quanto riguarda i non-custodial wallet, solamente il proprietario del wallet ha accesso alla sua chiave privata. I wallet forniti dalle piattaforme di exchange solitamente sono di tipo hot wallet. I secondi, che a loro volta si suddividono in hardware e paper wallet, non sono connessi ad Internet e sono entrambi non-custodial wallet. Vanno, però, connessi ad Internet ogni volta che si desidera effettuare una transazione.
  L'assenza di un quadro giuridico preciso determina l'impossibilità di attuare un'efficace tutela legale e contrattuale degli interessi degli utenti, che possono, pertanto, trovarsi esposti a dover subire ingenti perdite economiche, ad esempio in caso di condotte fraudolente, fallimento o cessazione di attività delle piattaforme exchange presso cui vengono custoditi i wallet. Di conseguenza, i cold wallet sono considerati una soluzione più sicura per conservare le proprie criptovalute, mentre gli hot wallet sono più pratici per coloro che effettuano transazioni quotidianamente.
- ICO: è l'acronimo di Initial Coin Offering e si tratta di un meccanismo finalizzato alla raccolta di fondi necessari a finanziare un progetto imprenditoriale basato sulla blockchain per aumentare il capitale, attirando risorse, oppure per creare e distribuire una nuova criptovaluta. Il processo delle ICO è simile a quello delle IPO (Initial Public Offering) e all'equity crowdfunding, ma a differenza di questi ultimi, l'ICO implica l'emissione di token digitali in luogo di strumenti finanziari tradizionali, come ad esempio le azioni. I token vengono offerti agli investitori che li acquistano con monete a corso legale oppure con criptovalute. La prima fase di una ICO consiste nell'ideare un modello di business e nell'esplorare le modalità di utilizzo della tecnologia blockchain. In un secondo momento, la realtà imprenditoriale che ha lanciato la ICO può richiedere un feedback anticipato da parte degli investitori. Successivamente, il nuovo progetto sviluppa e pubblica un white paper, ovvero un documento informativo che delinea le caratteristiche e le funzionalità della nuova criptovaluta, includendo anche informazioni sul ritorno dell'investimento realizzato da coloro che hanno finanziato il progetto. Infine viene attuato il progetto tramite l'utilizzo della blockchain coinvolgendo gli investitori.
  Il fenomeno delle ICO ha assunto dimensioni considerevoli, con una proliferazione massiccia di tale tecnica di finanziamento tra il 2017 e il 2018, quando sono stati raccolti tramite ICO oltre 11 miliardi di dollari, a causa della continua ricerca di mezzi di

finanziamento alternativi a quelli tradizionali e di mezzi di investimento capaci di garantire rendimenti significativi in tempi brevi. La mancanza di una regolamentazione specifica per tali operazioni e la possibilità per chiunque di lanciare una ICO, ha, però, permesso il verificarsi di numerose frodi.

2.4 PRINCIPALI CRIPTOVALUTE

Le cinque principali criptovalute, ordinate in base alla capitalizzazione di mercato sono:

- Bitcoin (BTC): è la criptovaluta più conosciuta, con prezzo e capitalizzazione di mercato maggiori rispetto a tutte le altre criptovalute, definite altcoin, proprio in quanto nate come alternative a Bitcoin. Il vantaggio principale di Bitcoin deriva dal fatto di essere stata la prima criptovaluta ad essere apparsa sul mercato, creando una base concettuale e tecnologica che ha successivamente ispirato lo sviluppo di migliaia di progetti concorrenti, dando vita ad un settore completamente nuovo. Bitcoin non ha un emittente centrale né un'autorità a cui sottostare, ma è governato dalla comunità dei suoi utenti, e il suo prezzo è definito solamente dalla domanda e dall'offerta di mercato. L'attività di mining per Bitcoin, basata sull'algoritmo sha-256, veniva remunerata, al momento del lancio nel 2009, con 50 bitcoin per ogni blocco validato, ma questo numero è soggetto al fenomeno dell'halving, ovvero ad un dimezzamento, ogni volta che vengono convalidati 210 mila blocchi, cosa che approssimativamente avviene ogni quattro anni. A maggio del 2020, è avvenuto il terzo halving che ha portato la remunerazione a 6,25 bitcoin. Inoltre, come per la maggior parte delle altre criptovalute, vi è una fornitura controllata, ovvero esiste un numero massimo (max supply) di bitcoin che possono essere messi in circolazione, che risulta pari a 21 milioni. Attualmente vi sono in circolazione 18,590,725 bitcoin. Secondo i calcoli, l'offerta di bitcoin raggiungerà la max supply nel 2140. Attualmente la capitalizzazione di mercato è di $ 577,812,263,189 e il prezzo, come si può vedere in figura 14, è pari a $ 31,035.43.

  Il Bitcoin è stato soggetto a molteplici interpretazioni nel tempo a causa del fatto che diverse entità abbiano avuto e abbiano tutt'ora diverse visioni del protocollo di Bitcoin, e con la sua evoluzione tecnologica e con il cambiamento della realtà circostante si sono susseguite ed accavallate diverse narrative. Nic Carter, il cofondatore di Coinmetrics.io, una piattaforma in cui sono riportati i dati relativi al mercato delle criptovalute, ha pubblicato il 29 luglio 2018 sulla piattaforma Medium un articolo dal titolo "Visions of Bitcoin", in cui espone le principali narrative sul Bitcoin. Esse sono:

  - E-cash proof of concept: rappresenta la narrativa iniziale, in quanto i primi crittografi e cypherpunks stavano valutando il progetto, determinando se potesse funzionare o meno. A causa del fallimento dei progetti di e-cash precedenti, è stato necessario che passasse del tempo prima che la gente si convincesse della fattibilità tecnica del progetto di Bitcoin e passasse a concezioni più elaborate del protocollo.

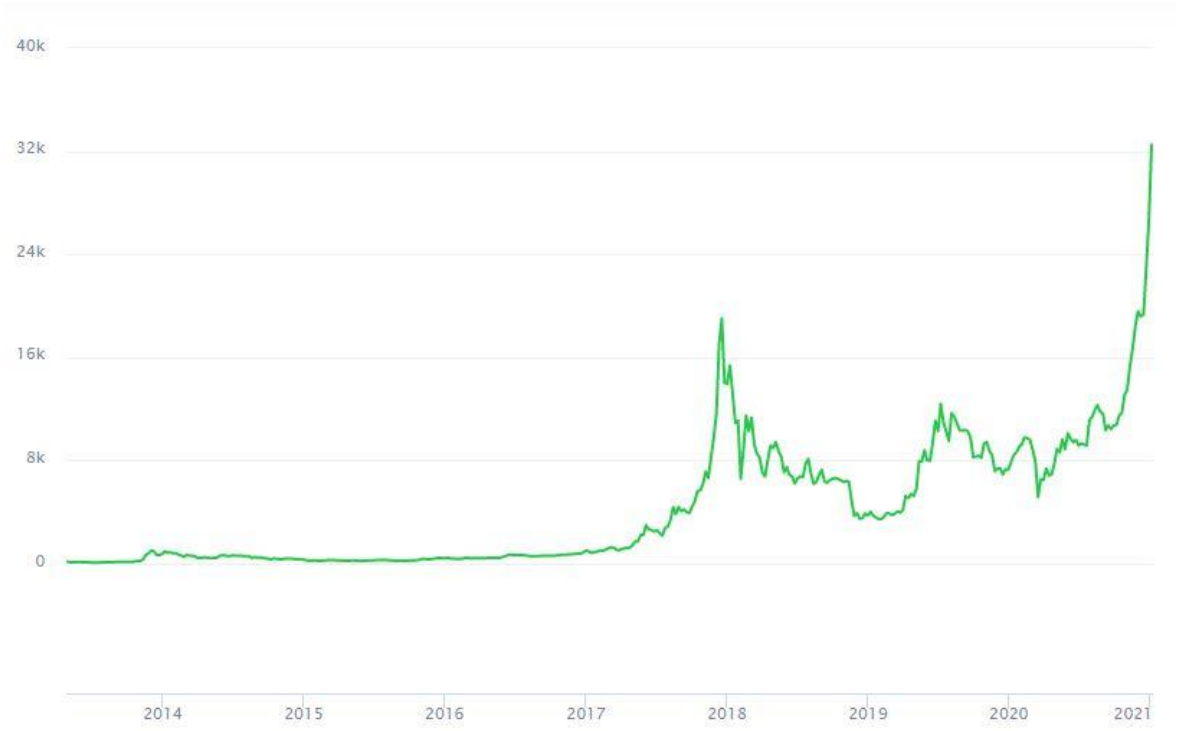

Figura 14: grafico Bitcoin (fonte: coinmarketcap.com, dati del 27/01/2021)

- Rete p2p per pagamenti economici: tale narrazione è stata una delle più popolari, in quanto dal whitepaper di Bitcoin sembrava proprio che questo fosse lo scopo a cui aspirava Satoshi Nakamoto.
- Oro digitale resistente alla censura: rappresenta il contraltare alla narrativa dei pagamenti p2p ed è il punto di vista secondo cui Bitcoin possa essere una riserva di valore non influenzabile dagli organismi istituzionali tradizionali.
- Valuta privata e anonima della darknet: rappresenta una narrativa secondo la quale Bitcoin sia utile per transazioni online anonime e in particolare per facilitare i traffici illegali del commercio online. Tale narrativa è in via d'estinzione poiché vi è stata una capillare diffusione di aziende per l'analisi dei dati on-chain, ovvero delle transazioni che avvengono sulla blockchain, che hanno permesso la comprensione del fatto che il protocollo del Bitcoin sia pseudonimo e non anonimo.
- Valuta di riserva per il mercato delle criptovalute: secondo tale narrativa bitcoin rappresenta la valuta di riferimento per il mercato delle criptovalute, la valuta in cui sono quotati i prezzi degli altri cryptoasset.
- Database condiviso programmabile: questa narrativa comporta la visione per cui Bitcoin possa essere un network per le transazioni di qualsiasi tipo di dati, oltre che per le transazioni di valuta, permettendo e facilitando molteplici casi d'uso.
- Attività finanziaria non correlata: tale narrativa approccia il bitcoin come una vera e propria attività finanziaria utile per la diversificazione del portafoglio, grazie alla sua bassa correlazione con altri asset tradizionali, sebbene ultimamente sia stata rilevata una correlazione inversa con il dollaro americano.

Nel grafico in figura 15 si può osservare come si siano susseguite le varie narrative e quale sia stata la loro popolarità.

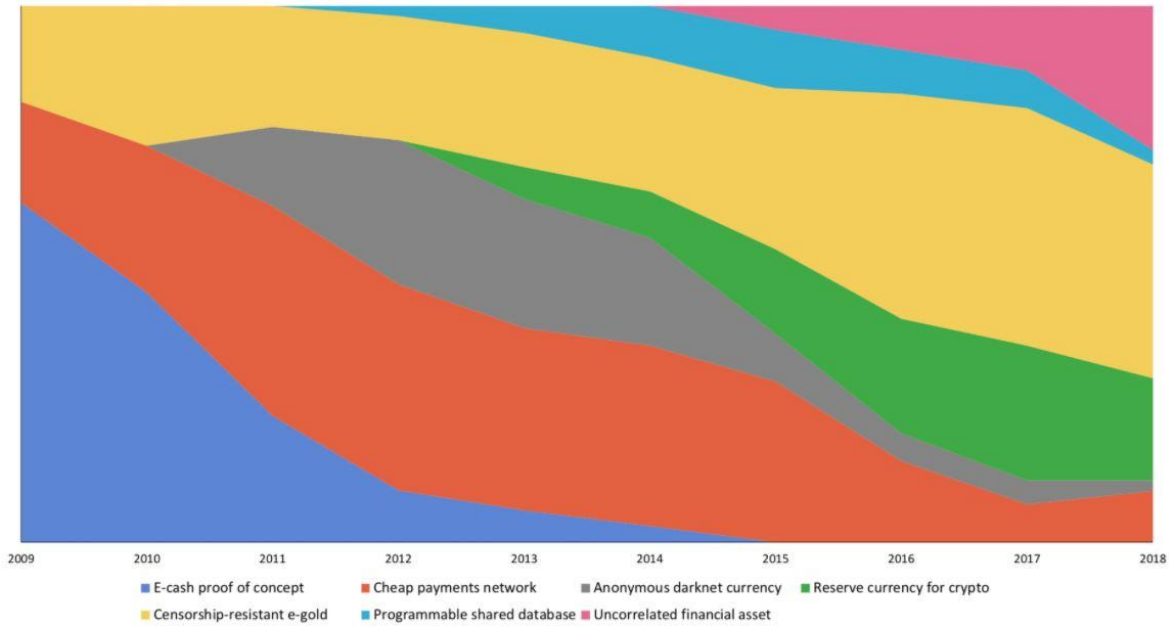

Figura 15: narrative del Bitcoin

- Ethereum (ETH): comunemente ci si riferisce ad Ethereum come una criptovaluta, ma, in realtà, Ethereum è una blockchain decentralizzata ed open-source che ha la propria criptovaluta chiamata Ether. Il white paper di Ethereum è stato pubblicato nel 2013 da Vitalik Buterin, insieme ad altri co-fondatori, e il progetto è stato lanciato ufficialmente il 30/07/2015.
  Ethereum è stato il pioniere del concetto di una piattaforma blockchain per smart contract, che consistono in protocolli informatici che eseguono automaticamente le azioni necessarie per adempiere ad un accordo tra più parti via Internet. Gli smart contract sono stati progettati per ridurre la necessità di intermediari tra i contraenti, diminuendo così i costi di transazione e aumentando al contempo l'affidabilità delle transazioni.
  Oltre agli smart contract, la blockchain Ethereum è in grado di ospitare altre criptovalute, definite anche token, in quanto non native di quella blockchain, attraverso l'uso del suo standard di compatibilità ERC-20. Sono numerose le criptovalute che si appoggiano alla blockchain Ethereum, molte delle quali sono tra le prime cento per capitalizzazione di mercato.
  L'obiettivo di Ethereum è quello di diventare una piattaforma globale per applicazioni decentralizzate, consentendo agli utenti di tutto il mondo di scrivere ed eseguire software resistenti alla censura, ai tempi di inattività e alle frodi.
  Una delle principali differenze tra l'economia della Bitcoin e quella di Ethereum è che quest'ultima non è deflazionistica, in quanto la sua offerta totale non è limitata. Gli sviluppatori di Ethereum giustificano questo fatto affermando che, non avendo una fornitura fissa, possono adeguare il tasso di emissione in modo tale da garantire sempre un'adeguata sicurezza.
  L'1/12/2020, come già esposto in precedenza, è avvenuto un aggiornamento rilevante per Ethereum, con il lancio della prima fase, che ha permesso il passaggio da un meccanismo di consenso basato sulla proof of work a uno basato sulla proof of stake.

Attualmente sono in circolazione 114,116,982 Ethereum, la capitalizzazione di mercato è di $ 115,386,054,275, e il prezzo, come si può vedere in figura 16 è pari a $ 1,006.87.

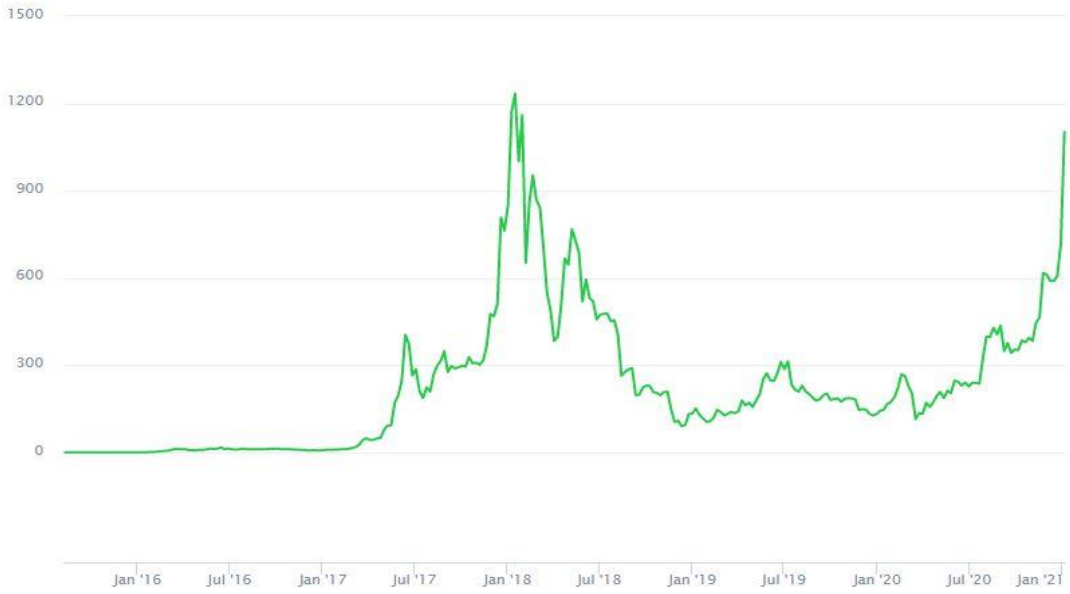

Figura 15: grafico Ethereum (fonte: coinmarketcap.com, dati del 27/01/2021)

- Theter (USDT): la USDT è una criptovaluta che si appoggia ad una blockchain esistente, ma più nello specifico è una stablecoin in quanto è ancorata al prezzo del dollaro statunitense con un rapporto 1:1. Il progetto è stato lanciato nel luglio 2014, con il nome di Realcoin, successivamente modificato in USTether e, infine, in USDT, da Brock Pierce, Reeve Collins and Craig Sellars, i quali hanno fondato la società Theter con sede a Hong Kong. Inizialmente, tale stablecoin si appoggiava alla blockchain di Bitcoin, ma, successivamente, il protocollo è stato aggiornato per permettere il funzionamento anche sulle blockchain di Ethereum, EOS, Tron, Algorand e OMG. L'ancoraggio di USDT al dollaro statunitense viene ottenuto dalla società mantenendo una riserva di dollari pari al numero di USDT in circolazione.
  L'obiettivo di USDT è quello di combinare la natura non vincolata delle criptovalute, che possono scambiate senza la presenza di un intermediario, con il valore del dollaro statunitense, che garantisce stabilità al prezzo dell'USDT. L'elevata volatilità che caratterizza il mercato delle criptovalute fa sì che il loro prezzo possa aumentare o diminuire in modo significativo all'interno di intervalli di tempo relativamente brevi, rendendole, dunque, inaffidabili come riserva di valore. L'USDT, invece, è protetto da queste fluttuazioni, diventando, perciò, un rifugio sicuro per coloro che investono in criptovalute durante periodi di alta volatilità. Inoltre, l'USDT fornisce un modo semplice per effettuare transazioni a livello globale in dollari statunitensi tramite blockchain, senza dover fare affidamento su un intermediario lento e costoso, come una banca o un fornitore di servizi finanziari.
  Il pregio della stabilità del valore, garantita dalla stablecoin, viene, però, controbilanciato dalla perdita di una completa decentralizzazione, a causa di come gli asset sottostanti debbano essere tenuti in riserva. Inoltre, proprio la garanzia che le stablecoin siano effettivamente ed adeguatamente coperte dall'asset sottostante rappresenta una grande sfida, che le società emittenti devono fronteggiare. Infatti, Tether ha dovuto affrontare

un'azione legale a New York a seguito di preoccupazioni secondo cui l'USDT poteva non essere coperto con un rapporto 1:1 con il dollaro statunitense.
Attualmente sono in circolazione 21,808,636,250 USDT, la capitalizzazione di mercato è di $ 21,805,758,773 e il prezzo, come si può vedere in figura 17, si aggira intorno a $ 1.00.

Figura 16: grafico USDT (fonte: coinmarketcap.com, dati del 27/01/2021)

- Ripple (XRP): XRP è la criptovaluta nativa del distributed ledger XRP Ledger, che viene usata sulla piattaforma di pagamento digitale RippleNet, gestita dalla società RippleLabs.
  L'idea alla base della piattaforma di pagamento, che mira a consentire transazioni istantanee a livello globale con qualunque tipo di valuta, è stata espressa per la prima volta nel 2004 da Ryan Fugger, ma il progetto è stato lanciato solamente il 26/09/2013 da Jed McCaleb e Chris Larson.
  A differenza della proof of work che avviene per Bitcoin, nell'XRP Ledger le transazioni vengono elaborate molto più velocemente, approssimativamente ogni 3-5 secondi, ogni volta che nodi validators indipendenti giungono ad un consenso sia sull'ordine che sulla validità delle transazioni XRP.
  Negli Stati Uniti, XRP è considerata alla stregua di un titolo, a differenza di Bitcoin ed Ethereum che sono considerate come commodities, in quanto, secondo i regolatori, XRP risulta conforme al test di Howey, secondo cui se un investimento monetario genera un profitto con l'impegno di terzi, allora l'investimento deve essere classificato come titolo, e perciò XRP deve sottostare alle federal securitis laws. Proprio a causa di una violazione di queste leggi, la Securities and Exchange Commission statunitense ha citato in giudizio RippleLabs, sostenendo che la società abbia venduto 1,3 miliardi di dollari di XRP e che il CEO Brad Garlinghouse e il co-fondatore Chris Larsen abbiano entrambi beneficiato personalmente della vendita di titoli digitali non registrati. La fase iniziale del processo comincerà il 22 febbraio 2021.

Per XRP vi è una max supply di 100,000,000,000 XRP, attualmente risultano in circolazione 45,404,028,640 XRP, la capitalizzazione di mercato è di $ 10,777,864,817 e il prezzo, come si può vedere in figura 18, è pari a $ 0.238319.

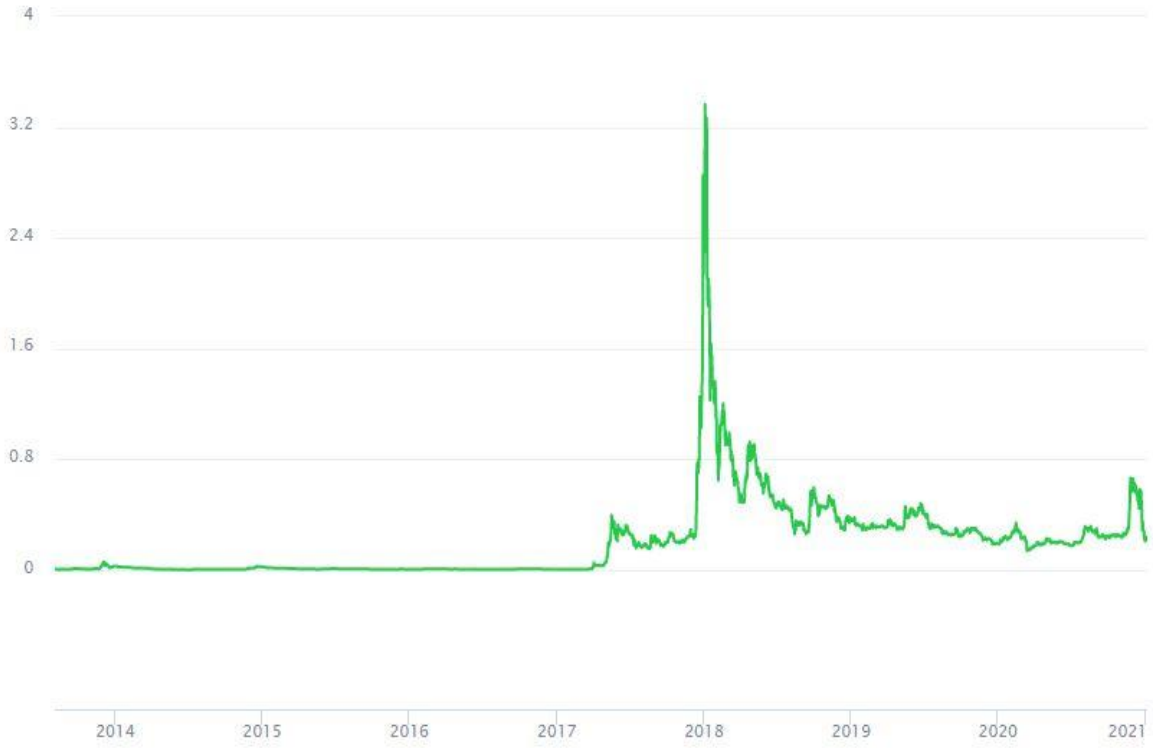

Figura 17: grafico XRP (fonte: coinmarketcap.com, dati del 27/01/2021)

- Litecoin (LTC): è una criptovaluta peer-to-peer creata da Charlie Lee, un ex dipendente di Google, nel 2011, sulla base del protocollo Bitcoin, ma differenziandosi per l'algoritmo di hashing utilizzato, la max supply, i tempi di transazione dei blocchi e alcuni altri fattori. Litecoin è stato rilasciato tramite un client open-source su GitHub il 7 ottobre 2011 e il Network Litecoin è entrato in funzione cinque giorni dopo, il 13 ottobre 2011.
  L'intenzione di Charlie Lee era di creare una versione più leggera di Bitcoin, dando la priorità alla velocità di conferma delle transazioni, che risulta essere di circa un paio di minuti per blocco. A differenza di Bitcoin, il mining di Litecoin viene effettuato tramite l'algoritmo scrypt al posto dello sha-256, permettendo anche ad hardware di basso livello come GPU o CPU di fare mining, ovvero rendendo tale attività maggiormente accessibile. Al momento del lancio, la remunerazione per ogni blocco minato era pari a 50 Litecoin, ma tale cifra ha subito un halving nel 2015 e continuerà a dimezzarsi fino a raggiungere la sua max supply pari a 84,000,000 Litecoin.
  Attualmente risultano in circolazione 66,241,368 Litecoin, la capitalizzazione di mercato è di $ 9,943,067,833 e il prezzo, come si può vedere in figura 19, è pari a $ 156.81.

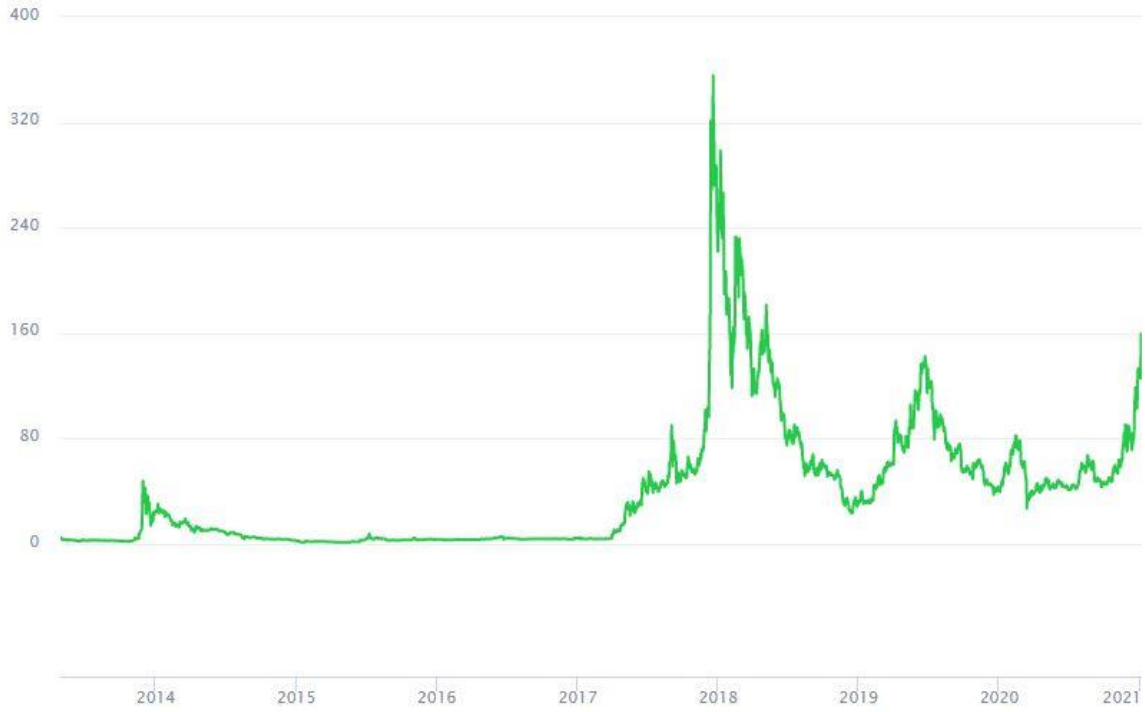

Figura 18: grafico Litecoin (fonte: coinmarketcap.com, dati del 27/01/2021)

2.5 CENTRAL BANK DIGITAL COIN

Le banche stanno cercando di trasformare la minaccia che arriva dalle criptovalute, le quali permettono la gestione delle transazioni e dei pagamenti senza la necessità di un intermediario, in una grande opportunità. Un esempio emblematico arriva dalle banche centrali, le principali istituzioni del mondo bancario che, come ha chiaramente messo in evidenza il report, del marzo 2019, del World Economic Forum, dal titolo "Central Banks and Distributed Ledger Technology: How are Central Banks Exploring Blockchain Today?", stanno attivamente lavorando a progetti che possono trarre vantaggio dalla blockchain. Il report illustra i dieci ambiti in cui le banche centrali stanno effettuando esperimenti e ricerche basati sulle blockchain e sulle DLT, tecnologie che possono generare nuove possibilità e migliorare processi inefficienti. Tra questi dieci, il punto chiave, sul quale è importante concentrare l'attenzione, riguarda la creazione e il sostegno a una Central Bank Digital Currency (CBDC).

La CBDC è una potenziale applicazione della blockchain e della DLT, e rappresenta una valuta strutturata come la versione digitale di una moneta a corso legale con il supporto delle banche centrali.

Dunque, la CBDC avrebbe caratteristiche diverse dalle criptovalute, in quanto rimarrebbe pienamente nell'orbita del sistema finanziario tradizionale, essendo gestita e controllata dalla banca centrale e collegata alla valuta fiat. Infatti, le CBDC non sono in conflitto con l'attuale sistema bancario, ma rappresentano uno strumento monetario innovativo in formato digitale valido come strumento di pagamento, come riserva di valore e come strumento progettato per aumentare la

sicurezza dal punto di vista dell'identificazione, della tracciabilità, della prevenzione, promuovere l'inclusione finanziaria, rendere i pagamenti più efficienti e ridurre i rischi di contraffazione e di frode.

Dal report emerge una classificazione delle CBDC, come si può notare nella rappresentazione denominata "the Money Flower" in figura 20, che vengono distinte in due grandi categorie:

- Wholesale CBDC: sono CBDC disponibili solo per le banche commerciali e le clearing houses, affinché possano effettuare pagamenti interbancari più efficienti e al di fuori dei tradizionali sistemi di pagamento.
- Retail CBDC: sono CBDC emesse per uso generale da parte dei consumatori, affinché possano effettuare pagamenti peer-to-peer e pagamenti tra consumatori e commercianti. Attualmente le banche centrali offrono pagamenti digitali solo alle banche e ad altri istituti finanziari, mentre la loro interazione con i consumatori è limitata al denaro contante. Una retail CBDC offrirebbe a tutti l'opportunità di accedere a una versione digitale della moneta emessa dalla banca centrale.

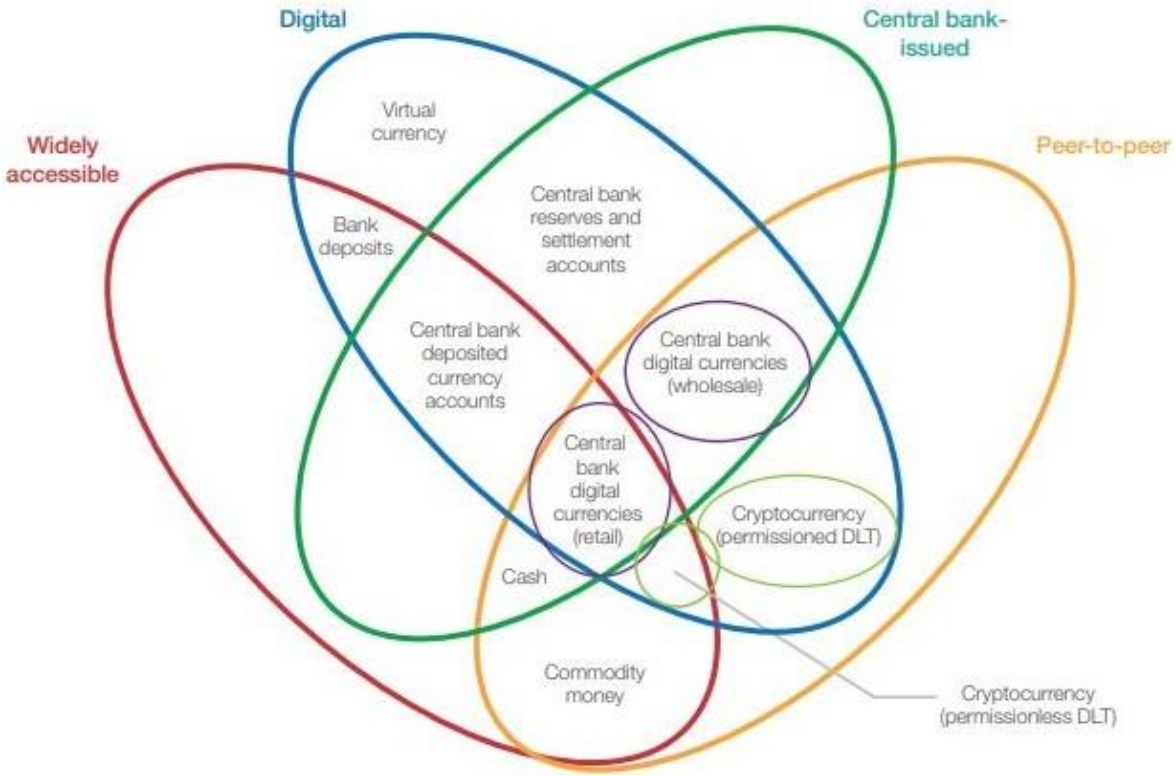

Figura 20: the Money Flower

Inoltre, come esplicitato nel report, i progetti delle banche centrali differiscono l'uno dall'altro sotto numerosi aspetti, il principale dei quali riguarda l'utilizzo della blockchain piuttosto che le tradizionali tecnologie centralizzate. Questo aspetto, che si rivela un punto critico su cui ogni progetto riflette attentamente, deriva dalla priorità che viene posta sul tema della privacy dei cittadini piuttosto che sull'efficace tracciamento delle transazioni. Nonostante ciò, appare evidente che la banca centrale, nel caso in cui optasse per l'utilizzo della tecnologia DLT, si affiderebbe ad una blockchain di tipo permissioned, in presenza della quale ai partecipanti deve essere consentito l'accesso alla rete da parte dell'autorità centrale al fine di visualizzare lo storico delle transazioni.

Gli analisti hanno studiato attentamente possibili strutture ed architetture per le CBDC, cercando di comprendere interamente le loro potenziali conseguenze sulla struttura giuridica dei crediti nei confronti della banca centrale e delle istituzioni private, ma, a prescindere dalla struttura delle CBDC, il loro emittente e fornitore finale è la banca centrale, che ne detiene la gestione e il controllo.
Nel paper intitolato "Impending arrival – a sequel to the survey on central bank digital currency", rilasciato a gennaio del 2020, la Bank for Internation Settlements (BIS) ha stimato che una significativa maggioranza, circa l'80%, delle banche centrali di tutto il mondo risulta impegnata nella ricerca sullo sviluppo della propria CBDC. Però, di questa percentuale, solo il 10% ha riferito di voler emettere una CBDC nel breve termine.

2.5.1 CBDC NEGLI STATI UNITI

Durante l'evento "Innovation Office Hours", avvenuto il 13/08/2020 presso la Federal Reserve Bank (FED) di San Francisco, Lael Brainard, membro del consiglio dei governatori della FED, ha dichiarato che la banca centrale statunitense sta testando le DLT per studiare gli effetti che una moneta digitale potrebbe comportare in relazione all'ecosistema dei pagamenti esistente, alla politica monetaria, alla stabilità finanziaria e al settore bancario. Ha, inoltre, esposto una panoramica dei molteplici esperimenti in corso presso il Board Technology Lab della banca centrale, affermando che è attiva una collaborazione tra la Federal Reserve Bank di Boston e i ricercatori del Massachusetts Institute of Technology (MIT) per dar vita ad un'ipotetica CBDC. Lael Brainard ha, tuttavia, affermato che i risultati di questa collaborazione sono destinati al conseguimento di una conoscenza più approfondita delle DLT e delle implicazioni apportate dalla presenza di una CBDC, piuttosto che ad offrire un prototipo per una futura moneta digitale emessa dalla FED.
Questa linea di pensiero è stata ricalcata successivamente, il 19/10/2020, durante il meeting annuale del Fondo Monetario Internazionale, da Jerome Powell, presidente della FED, il quale ha affermato che, per quanto riguarda un argomento così delicato, è meglio comprendere bene tutti gli aspetti connessi all'implementazione di una CBDC, piuttosto che essere i primi ad emetterla.
Nonostante la posizione della banca centrale, che comunque non si è rivelata contraria all'emissione di una CBDC, negli Stati Uniti è stata costituita da Christopher Giancarlo, ex presidente della United States Commodity Futures Trading Commission, una società no-profit, denominata Digital Dollar Foundation, che, insieme ad Accenture, ha pubblicato un'ambiziosa proposta, nell'ambito del Digital Dollar Project, per la realizzazione di ben nove programmi pilota per una CBDC statunitense. La Digital Dollar Foundation ha elencato una serie di componenti fondamentali che ritiene dovrebbero essere la linea di base per testare i modelli digitali del dollaro. La CBDC statunitense dovrebbe, dunque, essere innanzitutto emessa dalla FED, distribuita tramite l'esistente sistema bancario, tokenizzata in modo da essere un vero strumento al portatore, per cui chi detiene il token ne è anche il proprietario, accessibile a tutte le persone negli Stati Uniti, in grado di essere scambiata offline quando entrambe le parti sono in stretta prossimità fisica, ed inoltre dovrebbe garantire la privacy individuale e sostenere un processo di innovazione per quanto riguarda il settore dei pagamenti.
Come la maggior parte degli altri progetti relativa alla CBDC, la proposta mette in contrapposizione il diritto alla privacy con la necessità di una protezione normativa contro il riciclaggio di denaro

sporco. Le proposte pilota della Digital Dollar Foundation sono di ampia portata, ambiziose e sfaccettate, ma non sono state avviate dalla FED.

2.5.2 CBDC IN CINA

Per anni la Cina ha vietato progetti relativi alle criptovalute, ha messo al bando gli exchanges e ha persino bloccato i meetings che avevano le criptovalute come argomento, mentre cercava di capire come utilizzare la tecnologia dirompente della blockchian per una CBDC nazionale. La People's Bank of China ha lavorato al progetto di uno yuan digitale dal 2014 e, ad inizio 2020, ha lanciato l'iniziativa del Digital Currency Electronic Payment (DCEP) con un importante progetto pilota di implementazione in corso, diventando così la prima grande economia globale a lanciare una CBDC. Ad accelerare lo sviluppo della CBDC hanno contribuito, da una parte, la necessità di svincolare il renminbi dai sistemi di pagamento americani, Swift e Chips, che gestiscono la commercializzazione internazionale della valuta cinese, e, dall'altra, la volontà di battere sul tempo Libra (ora Diem), la moneta digitale a cui sta lavorando Facebook, che continua a suscitare dubbi e critiche da parte delle autorità internazionali.

Zhou Xiaochuan, presidente della Chinese Finance Association ed ex governatore della People's Bank of China, ha spiegato, durante la conferenza Eurasia Forum del 27/10/2020, che la strategia della banca centrale nella creazione di una valuta digitale si differenzia dall'orientamento dei paesi del G7, i quali rivolgo la loro attenzione alle sfide poste alla loro sovranità monetaria da Diem, Bitcoin e simili valute digitali, mentre la banca centrale cinese si sta focalizzando sull'uso della sua valuta digitale per i pagamenti retail domestici e per impedire al dollaro statunitense di diventare il mezzo di scambio più comune nel paese.

Sono state intraprese diverse iniziative per promuovere il DCEP, o yuan digitale, ed è stata aggiunta una rete blockchain di servizi per le imprese. Nel corso del 2020, il DCEP è stato distribuito a tutte le banche commerciali affiliate alla banca centrale cinese, ovvero alle quattro banche statali, Industrial and Commercial Bank of China, Agriculture Bank of China, Bank of China e China Construction Bank, le quali distribuiranno poi il DCEP ai consumatori. La People's Bank of China permette alle banche statali di testare diverse modalità di emissione e di distribuzione ai consumatori, così da poter individuare il modo migliore.

L'uso del DCEP è in corso di sperimentazione, con progetti pilota già operativi da aprile 2020, nelle principali città e nelle principali regioni economiche, come Shenzhen, Suzhou, Chengdu, Xiong'an, attraverso la distribuzione di DCEP sui wallet digitali di un numero definito di cittadini selezionati casualmente tramite delle estrazioni. I progetti pilota sono stati estesi anche a Pechino, alla regione del delta del fiume Yangtze, alle province di Tianjin, Hebei e Guangdong, alle città di Hong Kong e Macao e in futuro potrebbero estendersi alla Cina centrale e occidentale, con l'obiettivo di avviare un test a livello nazionale entro i giochi olimpici invernali del 2022.

Durante la conferenza Sibos 2020 del 5/10/2020, Fan Yifei, vice governatore della People's Bank of China, ha riportato i risultati dei primi progetti pilota effettuati tra aprile ed agosto 2020, affermando che la banca ha aperto 113.300 wallet digitali per i consumatori e 8.859 portafogli digitali aziendali, e che tali wallet hanno effettuato 3,1 milioni di transazioni di DCEP per un valore totale di 1,1 miliardi di renminbi (162 milioni di dollari), rendendo così il DCEP la CBDC più utilizzata in un contesto commerciale.

2.5.3 CBDC IN EUROPA

La Banca Centrale Europea (BCE) ha accelerato il ritmo degli studi condotti sulla realizzazione di una possibile CBDC durante la pandemia di coronavirus, e ad ottobre del 2020 ha pubblicato il "Report on a digital euro", che ha delineato gli scenari e i possibili requisiti per un futuro euro digitale.
Con un euro digitale si introdurrebbe una moneta elettronica, universalmente accettata, sicura ed affidabile, emessa dalla banca centrale, non concepita per sostituire il contante, ma per affiancarlo, ed accessibile a tutti, sia cittadini che imprese, accrescendo, così, l'inclusione finanziaria. Un euro digitale renderebbe i pagamenti di tutti i giorni più rapidi, semplici e sicuri, potrebbe sostenere la digitalizzazione dell'economia europea e incoraggiare attivamente l'innovazione nei pagamenti al dettaglio. Inoltre, contribuirebbe ad attutire l'impatto di eventi estremi, come calamità naturali o pandemie, quando i tradizionali servizi di pagamento sono inutilizzabili, ed avrebbe anche un ruolo cruciale se si diffondessero mezzi di pagamento digitali esteri, suscettibili di compromettere la stabilità finanziaria e la sovranità monetaria nell'area dell'euro. Le principali aree di riflessione individuate nel rapporto riguardano la privacy e il fatto che, essendo soggetto a tutte le norme che regolano le altre tipologie di pagamento, anche l'euro digitale dovrà rispettare le regole sul contrasto al riciclaggio, al finanziamento del terrorismo e all'evasione fiscale.
Nel rapporto vengono delineate due possibili tipologie di euro digitale, che potrebbero anche coesistere: una potrebbe essere utilizzata offline e, dunque, senza la necessaria presenza di un intermediario per l'elaborazione di ogni transazione; un'altra potrebbe essere utilizzata online, dove la presenza di un intermediario risulta indispensabile. La prima versione garantirebbe l'anonimato per gli utenti e dunque una migliore privacy dei dati, mentre la seconda offrirebbe maggiori possibilità di integrazione con i servizi esistenti. L'euro digitale per uso offline dovrebbe, ad un certo punto, essere gestito online per aggiungere o prelevare fondi dal dispositivo.
Inoltre, viene riferito che l'infrastruttura di fondo per la fornitura di un euro digitale potrebbe essere centralizzata, con tutte le operazioni registrate nel libro mastro della banca centrale, oppure prevedere un certo decentramento nei confronti degli utenti e/o degli intermediari controllati, ma che, indipendentemente dall'approccio adottato, l'infrastruttura sarebbe controllata in ultima analisi dalla banca centrale.
Durante il meeting annuale del Fondo Monetario Internazionale, il 12/10/2020, Christine Lagarde, presidente della BCE, ha affermato che la BCE deciderà se lanciare o meno un progetto sull'euro digitale verso la metà del 2021 e che per sviluppare tale progetto saranno necessari dai due ai quattro anni. Tra le principali preoccupazioni da parte della BCE vi è la necessità di garantire un alto livello di protezione dei dati e della privacy per chi utilizzerà la moneta digitale, e, su questo punto, come sottolinea Fabio Panetta, membro del comitato esecutivo della BCE e presidente dell'Eurosystem High-Level Task Force on Central Bank Digital Currency, organizzazione istituita dalla BCE nel 2020, una garanzia di tutela deriva dal fatto che la banca centrale non ha interessi commerciali legati ai consumi, diversamente dai fornitori privati di servizi di pagamenti.

3. ARTIFICIAL INTELLIGENCE, MACHINE LEARNING E DEEP LEARNING

Per comprendere nella sua totalità il concetto e il funzionamento di una rete neurale, che, mettendo in relazione gli argomenti trattati nei due capitoli precedenti, sarà oggetto del presente lavoro di tesi nel prossimo capitolo, è necessario definire il quadro generale in cui esso si inserisce, illustrando, innanzitutto, tramite una breve panoramica, il più ampio concetto di Artificial Intelligence (Intelligenza Artificiale o AI), che racchiude al suo interno il Machine Learning (Apprendimento Automatico) e il Deep Learning (Apprendimento Profondo o Apprendimento Gerarchico).

L'Artificial Intelligence, il Machine Learning e il Deep Learning sono termini che vengono spesso accomunati e utilizzati in modo intercambiabile, ma, nonostante si riferiscano allo stesso settore, si differenziano l'uno dall'altro, in quanto descrivono un insieme di innovative tecniche informatiche sempre più specifico. Dunque, è possibile interpretare tali termini alla luce della teoria matematica dell'insiemistica, definendo l'AI come l'insieme che comprende il Machine Learning, il quale, a sua volta ha come sottoinsieme il Deep Learning, come si può vedere in figura 21.

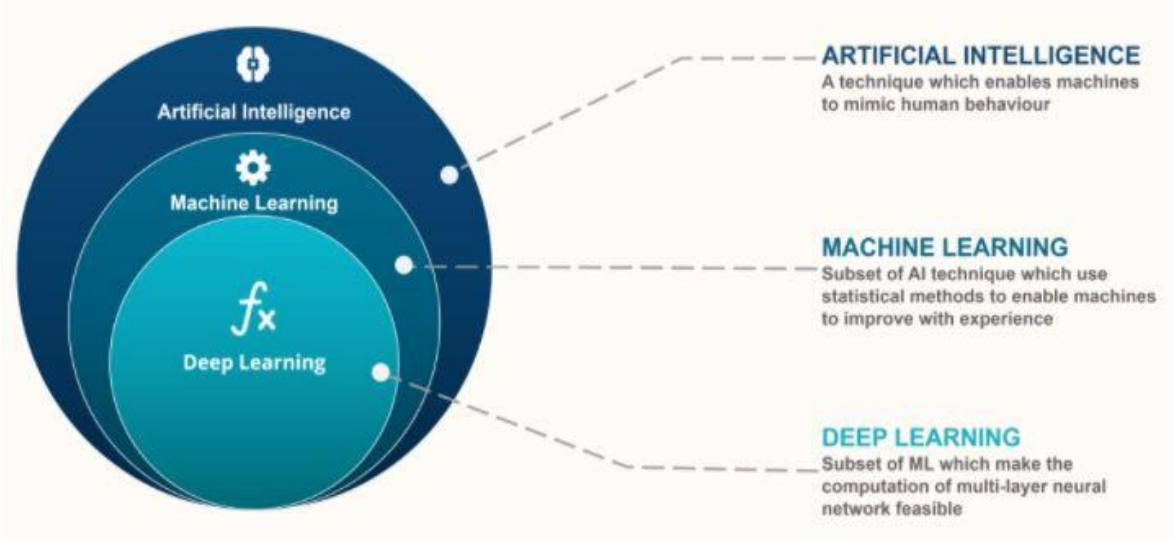

Figura 21: rappresentazione insiemistica delle tecniche di AI, ML e DL

L'AI rappresenta tutte quelle tecniche che permettono alle macchine, intese come dispositivi elettronici, di eseguire operazioni nello stesso modo in cui verrebbero compiute dall'intelligenza umana, garantendo maggiore efficienza e rapidità nell'esecuzione del compito.

Il Machine Learning consiste in un insieme di algoritmi e di metodi di programmazione, che, grazie all'introduzione dell'auto-apprendimento per le macchine, rende possibile l'AI.

Infine, il Deep Learning rappresenta uno dei molteplici approcci relativi al ML, che prende ispirazione dal comportamento cellulare del cervello umano, cercando di simularne i processi.

3.1 ARTIFICIAL INTELLIGENCE

Per decenni, la ragione che ha spinto ininterrottamente l'uomo a migliorare ed implementare i propri studi e sperimentazioni riguardanti l'intelligenza artificiale si identifica nel desiderio di creare una connessione inscindibile tra automazione e ragionamento. I primi tentativi si basavano su algoritmi matematici che, con il tempo, si sono evoluti e sono divenuti sempre più complessi ed affidabili. La differenza principale tra un algoritmo di AI e un algoritmo tradizionale risiede nel modo

in cui vengono definite le regole e le istruzioni che li costituiscono: in un algoritmo tradizionale, uno sviluppatore stabilisce regole specifiche che associano un output ad ogni tipo di input che il software riceve; invece, gli algoritmi di AI sono progettati in modo tale che, una volta ricevuti determinati input, essi riescano autonomamente a definire il proprio sistema di regole, anziché averne alcune già predefinite dallo sviluppatore.

La data della nascita dell'intelligenza artificiale, ufficialmente riconosciuta e coincidente con il raggiungimento di un ragionevole sviluppo delle tecnologie disponibili, parallelamente alle basi teoriche, è il 1956, anno in cui si tenne, presso il Dartmouth College di Hannover nel New Hampshire, un famoso seminario, durante il quale la nuova disciplina venne fondata programmaticamente e identificata come Artificial Intelligence da Marvin Minsky, insieme a Claude Shannon, John McCarthy, Herbert Simon e Nathaniel Rochester. Questo team di economisti ed informatici sosteneva che, se ogni aspetto dell'apprendimento e delle capacità cognitive in generale potesse essere descritto accuratamente, allora sarebbe stato possibile progettare e realizzare una macchina in grado di svolgere determinate operazioni sulla base di tali descrizioni, agendo perciò come se si trattasse di un'intelligenza umana. In altre parole, paragonando la mente umana ad un programma, grazie al quale, offrendo stimoli al cervello umano, si genera un ragionamento, il quale, a sua volta, produce un determinato comportamento, lo scopo di questi studiosi era quello di riuscire a creare una macchina capace di simulare il ragionamento umano.

Partendo dall'analisi funzionale di peculiari azioni intellettive umane, sono state individuate alcune funzioni di base che un sistema di AI deve essere in grado di compiere, per essere definito tale: compiere una o più azioni autonomamente, risolvere un problema attraverso specifiche funzioni cognitive, pensare razionalmente e agire razionalmente, ovvero avviare un processo che conduca al miglior risultato atteso sulla base delle informazioni e delle conoscenze a propria disposizione.

Nel 1980, è stata effettuata una suddivisione del grande settore che viene ricoperto dall'AI, ad opera del filosofo statunitense John Searle, il quale, nell'articolo pubblicato sulla rivista scientifica Behavioral and Brain Sciences e intitolato "Minds, brains, and programs", afferma che è necessario distinguere due tipologie di AI, le quali hanno dato vita a due correnti di pensiero sostenitrici rispettivamente dell'una e dell'altra tipologia:

- Weak AI o ANI (Artificial Narrow Intelligence): definita anche come intelligenza artificiale debole, essa si identifica in sistemi tecnologici in grado di agire e ragionare come se possedessero un cervello, senza, però, raggiungere le effettive funzionalità cognitive e capacità intellettuali proprie dell'essere umano. Infatti, l'obiettivo dell'intelligenza artificiale debole non è quello di realizzare macchine che posseggano un'intelligenza umana, ma consiste nello sviluppare sistemi in grado effettuare ragionamenti che possano aiutare l'essere umano a risolvere problematiche al fine di compiere con successo operazioni complesse, come ad esempio l'analisi rapida di significative quantità di dati da cui estrarre informazioni rilevanti.

  L'ANI elabora le informazioni relative ad un problema specifico, applicandovi un modello di ragionamento scientifico, ovvero analizzando casi simili al fine di accumulare maggiore esperienza, confrontando tra loro i diversi eventi, verificando empiricamente le ipotesi, elaborando una serie di soluzioni e scegliendo quella più razionale e congrua, svolgendo, così, in modo efficiente il compito assegnato. Dunque, si tratta di un'intelligenza parziale, specializzata, ma adattabile alle molteplici situazioni in cui sia necessario il problem solving. Inoltre, l'intelligenza artificiale debole non cerca di comprendere i processi cognitivi dell'essere umano nella sua totalità, non pretende di definire i processi che crea come

processi mentali, non tenta di riprodurre l'autoconsapevolezza nella macchina: il software elaborato agisce come se fosse un soggetto intelligente, ma non è necessario che lo sia davvero, a condizione che ottenga risultati soddisfacenti nei compiti che è chiamato a svolgere.

Infine, una macchina, dotata di intelligenza artificiale debole, non è capace di pensare in maniera autonoma, in quanto, nonostante riesca ad eseguire in maniera ottimale le operazioni per cui è stata costituita, realizzando un'intelligenza simulata, ha, tuttavia, bisogno della presenza dell'uomo per funzionare.

Un esempio di macchina basata sull'ANI è Deep Blue, uno dei primi successi dell'intelligenza artificiale, che, dopo aver disputato diverse partite contro l'allora campione del mondo di scacchi, Garry Kasparov, riuscì a batterlo nella data emblematica dell'11/05/1997. Deep Blue era un computer progettato appositamente per giocare a scacchi, pur non essendone consapevole, programmato nel 1985 dall'informatico cinese Feng-hsiung Hsu e dal canadese Murray Campbell nell'ambito del loro progetto di laurea, originariamente denominato ChipTest, alla Carnegie Mellon University di Pittsburgh. I due studiosi vennero poi assunti dalla IBM Research nel 1989, con l'incarico di portare avanti il loro progetto. Deep Blue, grazie al perfezionamento derivato dall'utilizzo delle risorse messe a disposizione dall'IBM, divenne capace di analizzare rapidamente un numero considerevole di varianti: era in grado di valutare 200 milioni di posizioni al secondo, potendo, così, analizzare dalle sei alle otto mosse successive, se sulla scacchiera erano presenti molti pezzi, ma giungendo addirittura ad esaminare una profondità di venti mosse in presenza di pochi pezzi. Al termine della partita, lo stesso Kasparov affermò che, dall'altra parte della scacchiera, aveva percepito un nuovo tipo di intelligenza.

I sistemi di intelligenza artificiale debole sono in grado di eseguire un singolo compito o una gamma di compiti, in cui possono ottenere prestazioni migliori di quelle conseguite dagli esseri umani, ma risultano limitati da un insieme ristretto di vincoli e parametri, a causa dei quali falliscono quando vengono adoperati per risolvere problemi al di fuori della loro sfera di applicazione, rivelandosi incapaci di trasferire da un contesto ad un altro le conoscenze acquisite. Ad esempio, il software AlphaStar, sviluppato da DeepMind, laboratorio di ricerca sull'intelligenza artificiale di proprietà di Google, riesce ad ottenere performance migliori della quasi totalità dei giocatori che si affrontano nel popolare gioco online di strategia in tempo reale StarCraft 2, ma lo stesso software non è in grado di giocare ad altri giochi, come Warcraft o Command & Conquer, sebbene prevedano le medesime tecniche e modalità di gioco.

Nonostante questa limitazione, i sistemi di intelligenza artificiale debole hanno dimostrato la loro utilità in molteplici campi di applicazione della vita quotidiana, permettendo numerose innovazioni, tra cui è possibile annoverare gli assistenti virtuali come Siri di Apple, Alexa di Amazon e Cortana di Microsoft, che permettono il riconoscimento vocale, le auto a guida autonoma, i filtri anti-spam, che catalogano automaticamente, all'interno della posta elettronica, i messaggi ritenuti indesiderati, gli strumenti di monitoraggio dei social media per contenuti pericolosi, gli strumenti di mappatura e previsione delle malattie, e i sistemi di profilazione, che attraverso l'applicazione di particolari algoritmi, a seguito della raccolta di dati personali e dell'identificazione di interessi ed abitudini caratteristiche, permettono l'elaborazione di specifici modelli in grado di suggerire e consigliare contenuti di intrattenimento o di marketing, come ad esempio le ricerche da

fare su Internet, i video da guardare su Youtube, la musica da ascoltare su Spotify e i film da vedere su Netflix.

- Strong AI o AGI (Artificial General Intelligence): definita anche intelligenza artificiale forte, essa si identifica in sistemi tecnologici denominati esperti, i quali, senza emulare i processi di pensiero o le capacità cognitive ed intellettuali caratterizzanti l'essere umano, sono in grado di sviluppare una propria intelligenza in modo autonomo. Una macchina dotata di intelligenza artificiale forte non rappresenta più solamente uno strumento atto alla risoluzione di problemi, ma diventa essa stessa una mente, che pensa, comprende ed agisce in una maniera non distinguibile da quella umana. L'AGI si basa sulla logica matematica, sul ragionamento, sulla dimostrazione automatica del problema, sull'analisi del linguaggio che permette la comprensione delle espressioni linguistiche umane da parte della macchina ed, infine, sulla pianificazione tramite algoritmi.

  Ricercatori e scienziati, che si occupano di AI, sono alla costante ricerca di un modo per ottenere un'intelligenza artificiale forte, ma risulta un compito arduo da portare a termine, in quanto bisognerebbe rendere le macchine coscienti, capaci di applicare la conoscenza acquisita dall'esperienza in un determinato campo in un settore diverso da quello di partenza, e, in ultima analisi, capaci di discernere i bisogni, le emozioni, le credenze e i processi mentali di altri soggetti intelligenti. La fase iniziale, che costituisce un ostacolo allo sviluppo di tale intelligenza artificiale, consiste in una conoscenza più approfondita e completa riguardo le funzionalità del cervello umano. Vi sono numerosi tentativi che mirano ad ottenere sistemi di AGI, i più rilevanti dei quali consistono nella creazione di supercomputer, ovvero sistemi informatici progettati per ottenere capacità di elaborazione estremamente elevate. Uno dei tentativi maggiormente degni di nota è il Fujitsu-built K, il supercomputer più veloce secondo la classifica TOP500, resa nota al pubblico durante la International Supercomputing Conference del 2011. Questo supercomputer, realizzato dall'azienda giapponese di prodotti e servizi per l'information technology Fujitsu insieme al centro giapponese di ricerca scientifica RIKEN, ha impiegato 40 minuti per simulare un singolo secondo di attività neurale. Da allora, numerose altre aziende hanno focalizzato le proprie risorse nella costruzione di supercomputer, come ad esempio l'azienda statunitense NVIDIA, la quale, insieme al consorzio universitario italiano Cineca, ha annunciato il 16/10/2020 la realizzazione di Leonardo, il supercomputer AI più veloce al mondo, che sarà attivato nel corso del 2021.

3.2 MACHINE LEARNING

Lo diffusione e l'aumento di applicazioni di intelligenza artificiale, come quelle descritte nel paragrafo precedente, è in gran parte dovuto ai processi di Machine Learning, che vengono utilizzati per sviluppare gli algoritmi che rappresentano la base del funzionamento delle applicazioni. Il Machine Learning è costituito, dunque, da un insieme di modelli, processi e tecnologie che rendono possibile l'intelligenza artificiale.

Il termine fu coniato da Arthur Samuel, informatico statunitense e uno dei pionieri dell'intelligenza artificiale, il quale, nel 1959, definì il Machine Learning come quella branca dell'informatica che permette ad una macchina di imparare ad eseguire un compito senza che sia stata esplicitamente programmata per farlo. Il termine fu successivamente ripreso dall'informatico statunitense e

professore universitario Tom Mitchell, il quale scrisse un libro intitolato “Machine Learning” nel 1997 e pubblicò un articolo dal titolo “The Discipline of Machine Learning” nel luglio del 2006, in cui ne diede una definizione più formale, affermando che un programma per computer apprende e impara se le sue prestazioni, misurate rispetto a determinate attività, migliorano a seguito dell’esperienza acquisita.

Dunque, il Machine Learning consente ad un algoritmo di imparare dall’esperienza passata, ovvero dagli input iniziali ricevuti, tramite una fase di training o di addestramento, e di adattarsi, apprendendo autonomamente come migliorare le proprie prestazioni, ovvero modificando l’output derivante da un medesimo input ricevuto a distanza di tempo. Come affermato precedentemente, questa è la caratteristica principale che determina una netta differenziazione dagli algoritmi tradizionali e che permette di comprendere la portata rivoluzionaria dell’intelligenza artificiale. Il Machine Learning ridefinisce il problema della programmazione, che deve fronteggiare le implicazioni derivanti dall’elevato numero di variabili e dall’imprevedibilità degli input in ingresso, permettendo agli algoritmi di imparare dagli errori commessi durante la fase di training, affinché questi possano poi comprendere automaticamente come reagire agli input successivi, senza che vi sia un sistema programmato anticipatamente per delineare il comportamento da adottare in ogni situazione contingente.

Con l’avvento dei Big Data, ovvero la notevole disponibilità di dati, che possono, poi, attraverso un’attenta gestione, essere accumulati ed analizzati per estrapolare informazioni rilevanti, i modelli di Machine Learning hanno avuto l’incentivo per un ulteriore sviluppo, in quanto proprio un’ingente quantità di dati è necessaria affinché gli algoritmi apprendano e generino il proprio sistema di regole per adattarsi nel miglior modo possibile in base alle situazioni che si verificano ogni qual volta ricevono nuovi dati di input. Gli algoritmi in questione utilizzano una serie di tecniche di elaborazione rapida per identificare l’esistenza di pattern o relazioni tra i dati necessari ad eseguire il processo di apprendimento, al fine di prendere decisioni o di effettuare previsioni precise ed accurate sui dati che verranno elaborati in seguito.

Dunque, i processi di Machine Learning necessitano di una prima fase, durante la quale avviene il processo di apprendimento automatico. Esistono, però, diversi modelli di apprendimento automatico, che utilizzano varie tecniche algoritmiche, a seconda della natura dei dati di input e dei risultati desiderati. Tali metodi possono essere applicati separatamente, oppure, se lo si ritiene necessario, quando i dati risultano più complessi da categorizzare, possono essere combinati gli uni con gli altri al fine di ottenere una maggiore accuratezza. Esistono, dunque, quattro tipologie di apprendimento automatico:

- Supervised learning: il modello di apprendimento supervisionato prevede l’utilizzo di dati etichettati, ovvero coppie di dati costituiti da input a cui corrispondono determinati output prestabiliti e corretti. L’algoritmo viene addestrato sulla base di questi dati etichettati, grazie ai quali può confrontare gli output prodotti con gli output previsti, analizzare gli errori commessi e, imparando da essi, modificare di conseguenza il modello. Nell’apprendimento supervisionato, l’obiettivo consiste nell’approssimare una funzione di mappatura che collega gli input e gli output dei dati di addestramento, determinando le caratteristiche di somiglianza, differenza e le correlazioni logiche in generale che collegano i primi ai secondi, in un processo volto a regolare e a rendere sempre più accurata e corretta la funzione di mappatura. Dopo aver addestrato l’algoritmo, risulta possibile applicare ciò che è stato appreso durante l’apprendimento supervisionato a nuovi dati di input, partendo

dai quali, grazie alla funzione di mappatura ricavata, è possibile effettuare previsioni sui futuri output.

L'apprendimento supervisionato, con cui vengono generalmente trattati problemi riguardanti regressione e classificazione, richiede un numero minore di dati di training rispetto ad altri metodi di apprendimento automatico ed, inoltre, risulta più facile, a causa della natura dei dati sui quali si effettua l'addestramento. Invece, una delle complicazioni che questo modello comporta riguarda il pericolo di overfitting, per cui la funzione di mappatura acquisita durante l'addestramento potrebbe rivelarsi estremamente corretta, tanto da associare precisamente le coppie di dati iniziali di input e output, senza avere però la capacità di generalizzare, prevedendo, peciò, in modo distorto gli output che dovrebbero derivare da nuove osservazioni. In altre parole, il problema di over-fitting si verifica quando un modello risulta sovra-ottimizzato, poiché si adatta in maniera perfetta ai dati osservati, ottenendo performance eccellenti durante le simulazioni su dati storici, per poi subire un decadimento delle prestazioni sui nuovi dati in modo più o meno veloce e marcato.

In figura 22 è possibile vedere una rappresentazione grafica del modello di apprendimento automatico appena analizzato.

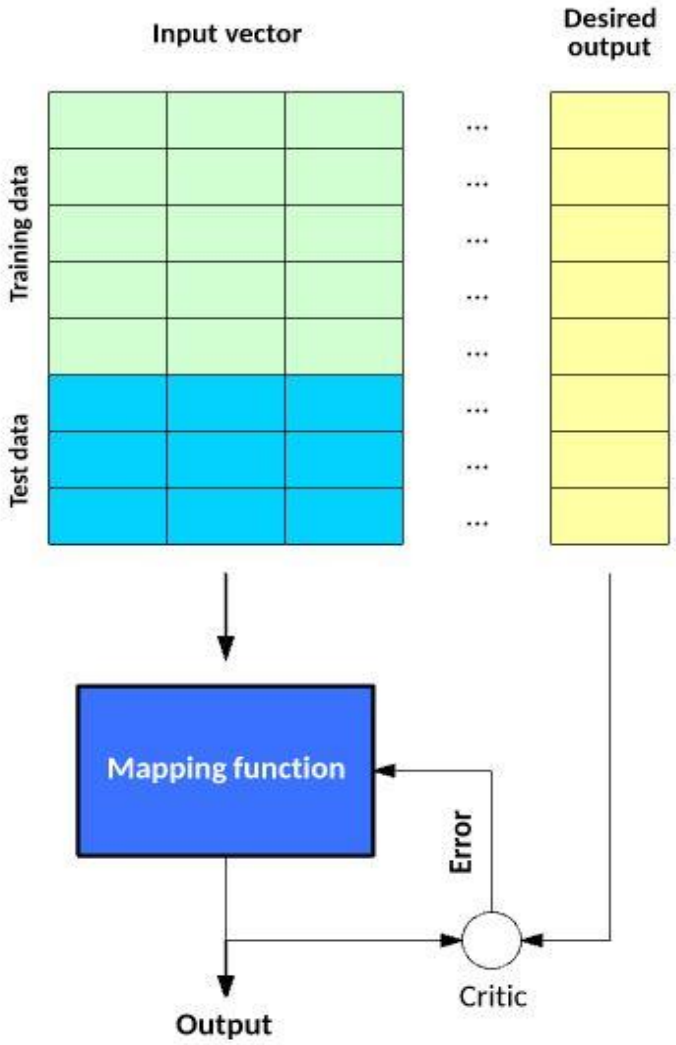

Figura 22: supervised learning

- Unsupervised learning: il modello di apprendimento non supervisionato prevede l'utilizzo di dati non etichettati, ovvero dati di input a cui non corrispondono output prestabiliti. A causa della natura dei dati di input, affinché un modello di apprendimento del genere funzioni sono necessarie ingenti quantità di dati. Infatti, durante l'apprendimento non supervisionato, l'algoritmo cerca di estrapolare le caratteristiche rilevanti nascoste nei dati, per identificare pattern e correlazioni esistenti tra questi, in modo tale da ordinarli e classificarli, ma, solamente in presenza di un numero consistente di dati, l'algoritmo sarà in grado di eseguire tale operazione. La funzione di mappatura non viene, dunque, migliorata sulla base di un rapporto esistente tra informazioni passate, ma viene dedotta tramite l'elaborazione dei dati e la ricerca di una struttura che permetta di descriverli e di utilizzarli per effettuare correttamente delle previsioni.

  L'apprendimento non supervisionato è modellato sul modo in cui gli esseri umani imparano, osservando ciò che li circonda: più aumenta l'esperienza, ovvero più sono disponibili esempi relativamente ad una determinata situazione, più diventa accurata la capacità di identificare e categorizzare quell'evento, anche e soprattutto sulla base di similitudini o differenze con altri eventi.

  L'apprendimento non supervisionato viene generalmente impiegato per l'esecuzione di due operazioni, ovvero il clustering, che riguarda la classificazione tramite un processo di raggruppamento di dati discreti in base alle caratteristiche riscontrate, e la dimensionality reduction, che ha l'obiettivo di trasformare i dati di input in caratteristiche rilevanti al fine dell'operazione da svolgere, riducendo, come si evince dalla denominazione della tecnica, la dimensionalità del set di dati, tramite un processo di eliminazione dei dati ritenuti non significativi.

  In figura 23 è possibile vedere una rappresentazione grafica del modello di apprendimento automatico appena descritto.
- Semi-supervised learning: l'apprendimento semi-supervisionato rappresenta una commistione delle due modalità di apprendimento analizzate in precedenza, in quanto nella fase di addestramento vengono utilizzati sia dati etichettati che non etichettati. Più specificatamente, l'apprendimento semi-supervisionato permette l'utilizzo di un set ristretto di dati etichettati, tramite i quali vengono individuate proprietà di correlazione che guidano la classificazione e l'estrazione delle caratteristiche rilevanti da un set più consistente di dati non etichettati, a cui vengono applicati i modelli derivati dall'iniziale apprendimento supervisionato. Dunque, i dati etichettati sono necessari per dare avvio al processo e possono migliorare considerevolmente la velocità e l'accuratezza dell'apprendimento.

  Poiché, spesso, i dati non sono immediatamente disponibili in forma etichettata e poiché la raccolta e la preparazione delle caratteristiche e degli output target sono le fasi più dispendiose nella risoluzione di un problema di Machine Learning, l'apprendimento semi-supervisionato risulta essere una soluzione ottimale quando sono presenti grandi quantità di dati grezzi e non strutturati e quando, dunque, è impossibile realizzare un algoritmo di apprendimento completamente supervisionato.
- Reinforcement learning: l'apprendimento per rinforzo è stato considerato inizialmente come un caso particolare dell'apprendimento supervisionato, ma successivamente la letteratura ha cominciato a trattarlo come una tipologia di apprendimento automatico a se

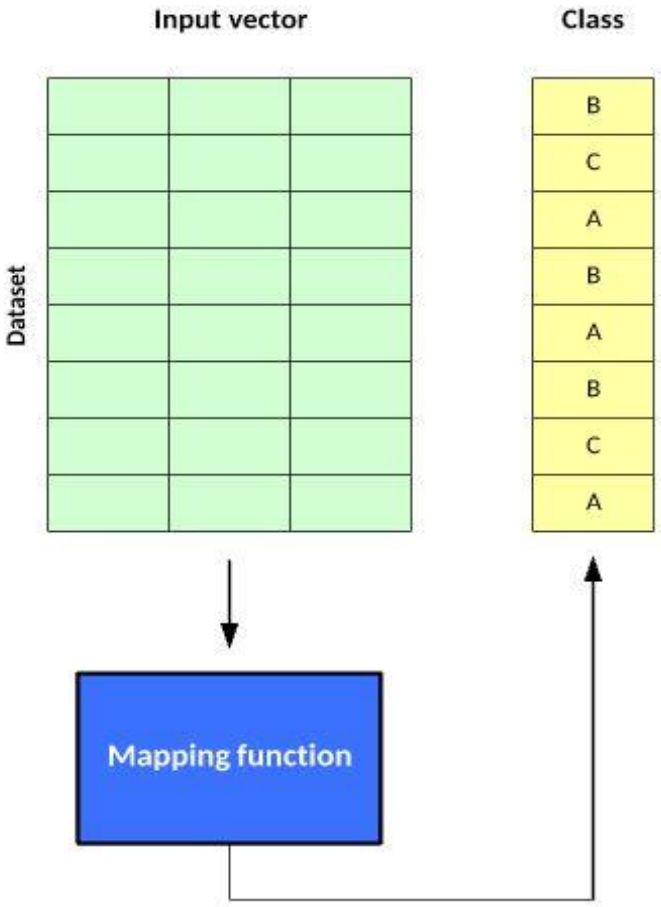

Figura23: unsupervised learning

stante, utile negli ambiti per i quali risulta inefficiente ed inefficace una codifica di uno specifico problema e la successiva elaborazione dello stesso tramite un algoritmo che offre un esplicito output per ogni possibile input. Infatti, se l'obiettivo o l'output target, che l'algoritmo deve provvedere a fornire, è determinato, fisso o binario, allora l'apprendimento può essere imposto tramite esempi, mentre, se il risultato desiderato è variabile e mutevole, allora è necessario un procedimento diverso. La peculiarità dell'apprendimento per rinforzo consiste nel feedback, denominato appunto segnale di rinforzo, che l'algoritmo riceve ogni qualvolta intraprende un'azione con la finalità di adattarsi ai cambiamenti a cui viene sottoposto. Se l'azione eseguita viene considerata come ammissibile, allora l'algoritmo riceverà un feedback positivo. Con questo procedimento, tramite un susseguirsi di tentativi ed errori, e mantenendo la memoria degli stati passati e dei feedback ricevuti, l'algoritmo apprende il metodo migliore per adattarsi alle diverse situazioni che si presentano, ovvero per intraprendere decisioni future ed effettuare previsioni il più accuratamente possibile.

Anche per quanto riguarda questa tipologia di apprendimento, è possibile fare riferimento all'approccio comportamentale adottato dagli esseri umani, i quali, associando una scelta corretta ad una ricompensa, cercheranno sempre di intraprendere la decisione giusta. Questo processo garantisce il raggiungimento del miglior risultato possibile, aumentando la probabilità che le scelte corrette si ripetano e riducendo contemporaneamente la probabilità che un errore si verifichi nuovamente.

In figura 24 è possibile vedere una rappresentazione grafica del modello di apprendimento automatico appena analizzato.

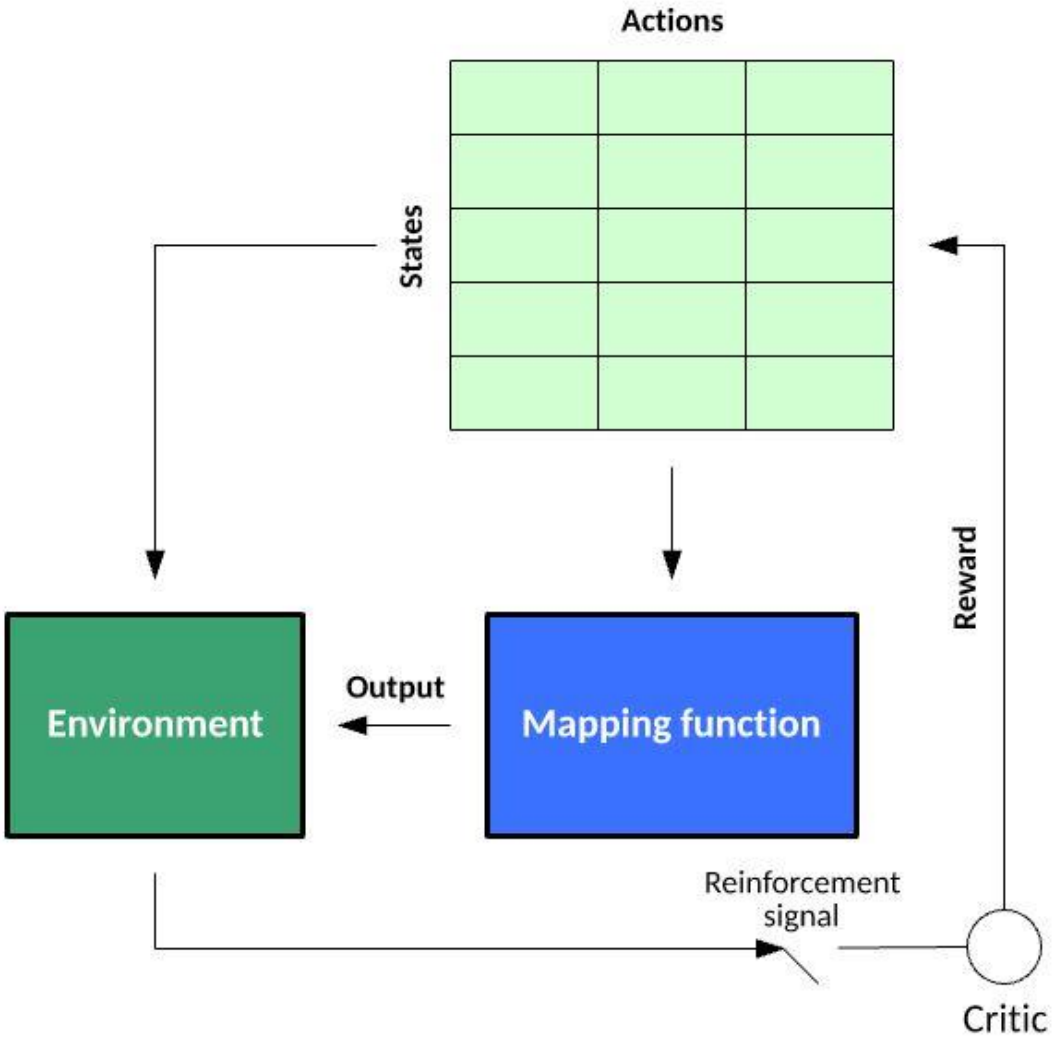

Figura 24: reinforcement learning

Inoltre, in aggiunta alla classificazione appena operata, è possibile annoverare diverse metodologie di machine learning, alcune delle quali sono applicabili solamente attraverso una delle tipologie di apprendimento descritte, mentre altre possono essere utilizzate sia con una determinata tipologia che con un'altra. Tra i metodi più noti vi sono il k-Nearest Neighbor, i Naive Bayes, i Support Vector Machine, i Random Forest, la regressione logistica e le reti neurali artificiali.

3.3 DEEP LEARNING E RETI NEURALI

Il deep learning, come visto in precedenza, è un sottoinsieme dei metodi di machine learning che prende ispirazione dal funzionamento del cervello umano nell'elaborazione dei dati e nella creazione di pattern da utilizzare nel processo decisionale. Il deep learning, a sua volta, è costituito da un insieme di modelli, i quali, proprio per il fatto di imitare la struttura delle reti neurali cerebrali, vengono denominati reti neurali artificiali o artificial neural network (ANN), che hanno la capacità di creare un sistema rappresentativo delle molteplici relazioni esistenti tra le variabili causali di un

sistema complesso, manifestando, di conseguenza, elevata attitudine ad esprimere con regolarità valutazioni appropriate e comportamenti adeguati in presenza di situazioni apparentemente caotiche. In questo caso, dunque, il processo di apprendimento viene definito profondo poiché la struttura delle reti neurali artificiali può essere costituita da più livelli, denominati layer, ognuno dei quali può contenere molteplici unità di elaborazione che trasformano i dati di input in informazioni che verranno utilizzate dal livello successivo per un determinato compito predittivo. Grazie a questa struttura, l'algoritmo che permette il funzionamento della rete neurale impara attraverso la propria elaborazione dei dati, in un processo che non segue un andamento lineare.

Il deep learning è il risultato di studi la cui origine risale agli inizi degli anni '40 del secolo scorso, ma appare come un'innovazione più recente in quanto, per un determinato periodo di tempo, verso la fine del XX secolo, il suo sviluppo è stato limitato a causa dell'evoluzione di altri modelli di machine learning, ed inoltre poiché, riflettendo le prospettive dei diversi ricercatori che si sono cimentati in tale materia, ha assunto diverse denominazioni durante gli anni. Infatti, come affermano Ian Goodfellow, Yoshua Bengio e Aaron Courville, nel loro libro intitolato "Deep Learning" e pubblicato nel 2016, è possibile suddividere lo sviluppo del deep learning in tre fasi: la prima, che va dal 1940 al 1960, durante la quale il deep learning veniva definito come cibernetica, la seconda, che comprende gli anni tra il 1980 e il 1995, quando il deep learning era noto come connessionismo, ed infine la terza, a partire dal 2006, in cui viene utilizzato il termine deep learning.

Nel 1943 è stato proposto il primo modello di neurone biologico dal neurofisiologo Warren McCulloch e dal matematico Walter Pitts nel saggio intitolato "A Logical Calculus of the Ideas Immanent in Nervous Activity", pubblicato sulla rivista Bulletin of Mathematical Biophysics. Tale modello computazionale, che aveva l'obiettivo di riconoscere due differenti categorie di input, verificando se l'output, definito da una funzione che ponderava gli input tramite l'attribuzione manuale di pesi, fosse positivo o negativo, ha segnato il punto di partenza per la ricerca sulla teorie cognitive del cervello e per la loro applicazione in materia di intelligenza artificiale.

Nel 1958, l'idea di MCCulloch e Pitts è stata recuperata e ampliata dallo psicologo statunitense Frank Rosenblatt, il quale, nel suo articolo intitolato "The perceptron: A probabilistic model for information storage and organization in the brain", pubblicato sulla rivista Psychological Review, ha proposto l'idea di neurone artificiale, definito perceptron, che si presentava come un semplice classificatore lineare binario in grado di apprendere efficacemente la regola necessaria per riconoscere due classi di input diverse e linearmente separabili. Oltre alla realizzazione fisica del perceptron, tramite una macchina elettromeccanica costituita da potenziometri, motori e fotocellule, l'innovazione conseguita dal modello di Rosenblatt risulta essere proprio l'applicazione di un semplice algoritmo di addestramento, grazie al quale il perceptron riusciva a modificare automaticamente i pesi degli input.

Due anni più tardi, un modello simile a quello di Rosenblatt venne sviluppato da Bernard Widrow, professore della Stanford University, e dal suo studente Marcian Edward Hoff, i quali idearono un macchinario denominato ADALINE (Adaptive Linear Element), che, a differenza del perceptron, utilizzava la discesa stocastica del gradiente (stochastic gradient descent) quale algoritmo di apprendimento per adattare i pesi con cui modificare la funzione di output.

L'entusiasmo che questi primi modelli di single-layer neural network avevano generato, portando alla nascita del settore della cibernetica, fu smorzato dalle opinioni esposte dai matematici ed informatici Marvin Minsky e Seymour Papert, i quali, nel 1969, pubblicarono un libro intitolato "Perceptrons", in cui espressero il loro scetticismo nei confronti delle reti neurali artificiali, cercando di deviare l'attenzione su altri metodi per lo sviluppo dell'intelligenza artificiale. Per

raggiungere il loro obiettivo, Minsky e Papert dimostrarono i limiti dei modelli di Rosenblatt e di Widrow e Hoff, affermando che questi, nonostante un opportuno addestramento, erano in grado di riconoscere solamente funzioni linearmente separabili, non riuscendo a risolvere problemi più complessi a causa della crescente complessità computazionale che l'addestramento avrebbe richiesto. Il risultato di tali critiche fu un significativo declino della ricerca e dei finanziamenti per progetti sulle reti neurali per un periodo di circa dieci anni.
La seconda ondata di sviluppo si è verificata a partire dagli anni '80, grazie a idee innovative e all'accesso ad una maggiore potenza computazionale, con cui è stato possibile un aumento esponenziale della ricerca sulle reti neurali. In tale periodo, nel contesto della scienza cognitiva, ovvero quell'approccio interdisciplinare allo studio della comprensione della mente, che combina diversi livelli di analisi, è sorto il movimento denominato connessionismo o processo distribuito parallelo, la cui idea fondamentale consisteva nel combinare un elevato numero di semplici unità di calcolo, come il perceptron, collegandole tra loro in un network, al fine di ottenere un comportamento intelligente della rete che queste unità costituiscono. Questa intuizione, applicabile sia ai neuroni del sistema nervoso biologico, sia alle unità presenti nei modelli computazionali, ha permesso lo sviluppo delle prime reti multistrato (multilayer neural networks), definite anche multilayer perceptrons (MLP).
Inoltre, durante il movimento del connessionismo sono stati definiti diversi concetti rilevanti che costituiscono le basi del deep learning per come è inteso attualmente. Uno dei concetti principali consiste nell'algoritmo di retro-propagazione (back-propagation), che, sebbene fosse stato sviluppato nel 1969 dal professore di ingegneria alla Stanford University Arthur Bryson e dal matematico Yu-Chi Ho, fu reso popolare nel 1986 dal team composto dallo psicologo David Rumelhart, dall'informatico Geoffrey Hinton e dal matematico Robert Williams. Infatti, Bryson e Ho idearono il metodo della back-propagation, esposto nel libro "Applied optimal control: optimization, estimation, and control", lavorando nel generico campo del controllo ottimale, ovvero una teoria matematica che, dopo aver determinato un algoritmo capace di ottimizzare una funzione di costo soggetta ad alcuni vincoli, applica tale metodo di ottimizzazione, derivante dal calcolo delle variazioni, a specifici ambiti di studio per definire le politiche ottimali di controllo dei sistemi dinamici che caratterizzano tali ambiti. Successivamente, nel 1986, quando Rumelhart, Hinton e Williams pubblicarono l'articolo "Learning representations by back-propagating errors" sulla rivista Nature, l'algoritmo di back-propagation venne applicato all'ambito delle reti neurali, al fine di permettere un migliore e più efficace processo di addestramento.
La seconda ondata di ricerche sulle reti neurali è durata fino alla metà degli anni '90, periodo in cui le società, focalizzate sullo sviluppo di reti neurali e di altri metodi di intelligenza artificiale e alla ricerca di finanziamenti per perfezionare i propri progetti, hanno iniziato a promettere risultati tanto ambiziosi quanto irrealistici, generando una profonda delusione negli investitori, i quali non vedevano soddisfatte le proprie aspettative. Inoltre, vi è stato contemporaneamente un progresso in altri metodi di machine learning, come ad esempio le kernel machines o i modelli grafici, entrambi inclusi nella categoria dei metodi di apprendimento Bayesiano. Questi due fattori hanno determinato un calo della popolarità delle reti neurali, che si è protratto fino al 2006.
Dal 2006 si è verificata una svolta nelle ricerche riguardanti le reti neurali, grazie al fondamentale contributo di Hinton, il quale faceva parte del Neural Computation and Adaptive Perception (NCAP), un programma di ricerca istituito dal Canadian Institute for Advanced Research (CIAR), che ha unito i molteplici gruppi di ricerca sul machine learning, guidati proprio da Hinton all'Università di Toronto, da Yoshua Bengio all'Università di Montreal e da Yann LeCun alla New York University,

contribuendo così a mantenere vivi gli studi in tale ambito. Proprio nel 2006, Hinton, assieme ad altri colleghi, ha pubblicato due articoli, intitolati rispettivamente “Reducing the dimensionality of data with neural networks” e “A fast learning algorithm for deep belief nets”, che hanno reso popolare il termine deep learning, per sottolineare che i ricercatori erano in grado di addestrare reti neurali più profonde di quanto non fosse stato possibile in precedenza, focalizzando, inoltre, l’attenzione sull’importanza teorica della profondità. Grazie agli studi condotti in questo periodo, alla crescita continua nella disponibilità di dati, al miglioramento delle infrastrutture hardware e software e dunque alla maggiore potenza computazionale, con l'introduzione delle unità di elaborazione grafica (GPU) e dell'elaborazione distribuita, è stato possibile addestrare modelli di reti neurali profonde di dimensioni sempre maggiori, permettendo l’ascesa delle reti neurali come algoritmo principale nell’apprendimento automatico, ampliando, così, il campo di applicazioni e garantendo un’elevata precisione nella risoluzione di problemi sempre più complessi.

3.3.1 MODELLO MATEMATICO DEL NEURONE ARTIFICIALE

Le reti neurali rappresentano il punto di contatto tra diverse discipline, quali la neurologia, la psicologia e l’intelligenza artificiale. Esse cercano di emulare, attraverso dispositivi software o hardware, i comportamenti delle trasmissioni sinaptiche che avvengono nel cervello durante la fase di apprendimento e di trasmissione dell’informazione e, dunque, la loro costituzione e il loro funzionamento si ispira all’elemento basilare tramite cui il cervello elabora le informazioni, ovvero il neurone.
Un neurone biologico, come si può vedere in figura 25, è caratterizzato da una semplice struttura, costituita da tre elementi essenziali:

- Soma: rappresenta il corpo cellulare in cui è presente il nucleo che dirige tutte le attività del neurone.
- Assone: singola fibra lunga che trasmette informazioni dal corpo cellulare ai dendriti di altri neuroni.
- Dendriti: fibre che ricevono informazioni da altri neuroni e inoltrano tali messaggi al corpo cellulare.

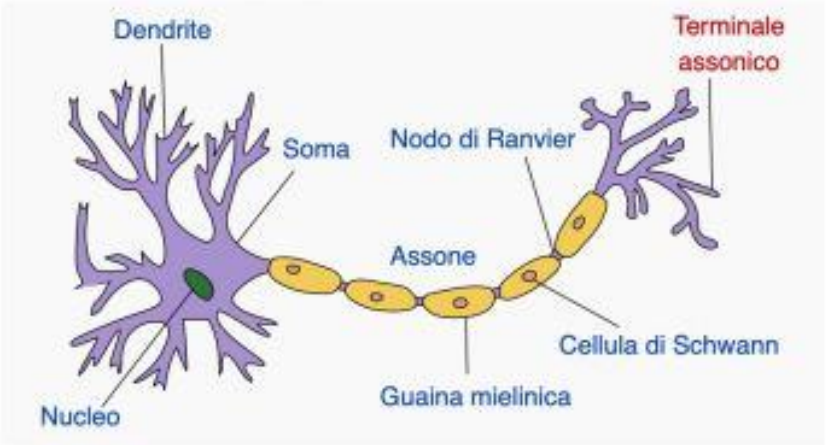

Figura 259: rappresentazione del neurone biologico

I neuroni ricevono input elettrici dai dendriti, assorbono energia e la rilasciano tramite gli assoni, che rappresentano i canali di output, ad altri neuroni vicini, trasmettendo l'informazione sotto forma di impulso elettrico attraverso strutture specializzate chiamate sinapsi. Nel caso in cui il segnale favorisca o meno l'attivazione del neurone ricevente, la sinapsi può essere eccitatoria o inibitoria e solo dopo un certo periodo di tempo, chiamato tempo refrattario, il neurone può generare un ulteriore impulso.
Analogamente, una rete neurale artificiale è costituita dall'interconnessione di semplici unità di elaborazione, i cosiddetti neuroni artificiali, definiti anche perceptron, dal modello di Rosenblatt, o più semplicemente nodi, che raccolgono e classificano le informazioni ricevute da input esterni e le trasmettono, dopo averle elaborate, alle unità con cui sono collegati, in base alla specifica architettura secondo cui sono disposti.
Il neurone può essere descritto matematicamente da

$$y_k = \varphi\left(\sum_{j=0}^{n} x_j w_{jk} + b_k\right)$$

oppure, più sinteticamente da

$$y_k = \varphi(v_k + b_k)$$

dove $x_1, x_2, \dots, x_n$ sono i segnali di input, b_k rappresenta il bias del neurone k, $w_1, w_2, \dots, w_n$ sono i pesi sinaptici, v_k è il potenziale d'azione, φ è la funzione d'attivazione e, infine, y_k è l'output del neurone. La rappresentazione grafica di tale modello matematico viene mostrata in figura 26.

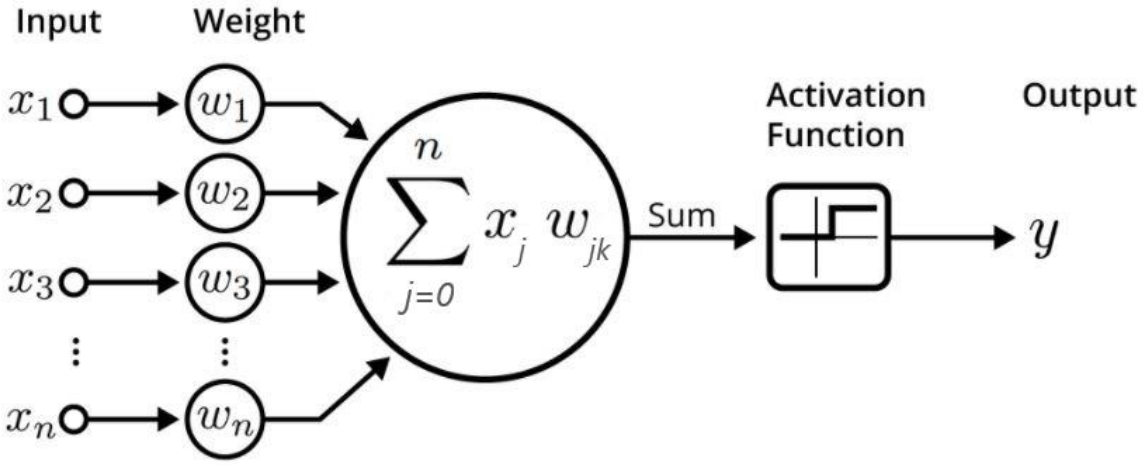

Figura 2610: rappresentazione grafica del modello matematico del neurone

Il neurone artificiale riceve in ingresso degli input, che possono derivare dall'esterno o da altri neuroni artificiali che compongono la rete, ognuno dei quali viene associato ad un peso corrispondente, denominato peso sinaptico. Più nello specifico, il segnale di ingresso x_j, dove j rappresenta le sinapsi connesse al neurone k, viene moltiplicato per il peso w_{jk}. I pesi sono parametri numerici che determinano l'intensità con cui ciascun input incide sull'operazione compiuta dal neurone, oppure, se gli input derivano da un altro nodo, allora i pesi rappresentano l'intensità con cui quel determinato neurone influenza quello successivo nella disposizione della

rete. Una volta ricevuti tutti gli input, il neurone effettua una somma ponderata dei segnali in ingresso moltiplicati per i corrispondenti pesi, e il risultato di tale operazione viene elaborato tramite la funzione di attivazione per calcolare il valore dell'output da trasmettere al neurone successivo o da utilizzare per l'obiettivo con cui è stata costituita la rete neurale. Oltre agli input è presente anche un bias, che, in modo simile all'intercetta della regressione lineare, contribuisce all'effetto della sommatoria, aumentando o diminuendo l'ingresso netto degli input ponderati nella funzione di attivazione.

Un neurone può essere rappresentato nello spazio come una linea infinita, e, dunque, i punti di intersezione delle linee derivanti dalla combinazione di più neuroni permettono la definizione di una frontiera di separazione tra le caratteristiche del problema in analisi: più neuroni sono presenti, più il modello di rete neurale è flessibile e adattabile alla risoluzione di differenti problemi. Il bias è necessario affinché le linee generate dai neuroni non siano vincolate a passare per l'origine degli assi cartesiani.

I pesi costituiscono l'elemento principale nelle reti neurali, in quanto permettono il processo di estrazione delle informazioni dai dati e la trasmissione di queste ultime tra i vari nodi, ed inoltre poiché la regolazione dei pesi rappresenta il metodo più utilizzato per l'addestramento dei modelli di deep learning.

Lo sviluppo avvenuto negli ultimi anni ha determinato il superamento del limite dimostrato dai primi modelli e rilevato da Minsky e Papert, per cui risultava possibile l'approssimazione delle sole funzioni lineari. Tale evoluzione negli studi compiuti ha permesso, così, alle reti neurali di poter approssimare qualsiasi tipo di funzione, grazie all'utilizzo di diverse funzioni di attivazione, dette anche non-linearity, ovvero funzioni matematiche che si occupano del processo di astrazione dei dati elaborati, convertendo l'input ricevuto dal neurone in output. Le funzioni di attivazione permettono di distinguere i modelli di deep learning dai modelli di regressione lineare, in cui si cerca semplicemente di approssimare la distribuzione dei dati con una retta. Nei neuroni biologici il potenziale d'azione viene trasmesso integralmente una volta che la differenza di potenziale tra le membrane supera una determinata soglia e lo stesso procedimento avviene anche per quanto riguarda i neuroni artificiali, il cui comportamento, inteso come output prodotto, viene, però, adeguato a seconda delle necessità, ed è determinato dalla funzione di attivazione adottata. Le funzioni di attivazione maggiormente conosciute e utilizzate sono:

- Funzione a gradino (Step function): tale funzione calcola un segnale di uscita diverso a seconda che il suo argomento si trovi al di sopra o al di sotto di una certa soglia. Viene utilizzata, dunque, nei casi in cui l'output assume il valore 0 se l'argomento della funzione è negativo e 1 se è positivo, come è possibile notare in figura 27. Dal punto di vista matematico, la funzione a gradino è riassumibile con

$$Step = \varphi(v) = \begin{cases} 1, & v \geq 0 \\ 0, & v < 0 \end{cases}$$

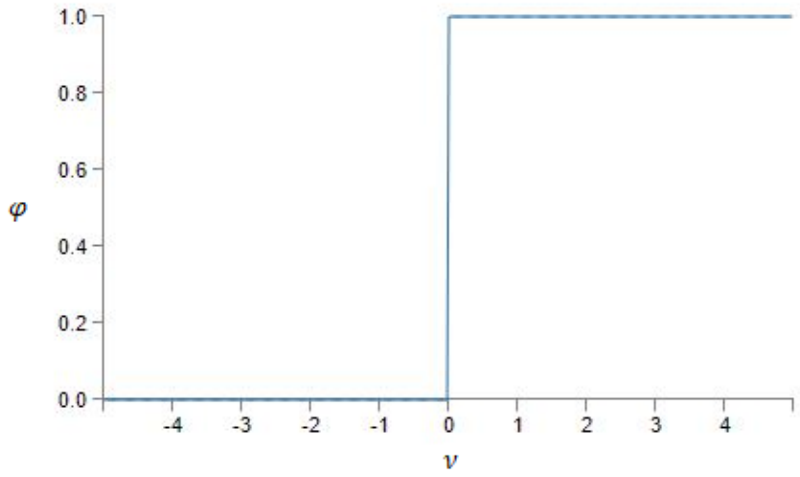

Figura 27: Step function

- Funzione sigmoide o sigmoidale: tale funzione trasforma l'argomento in un numero compreso nell'intervallo tra 0 e 1, e risulta centrata in 0.5 per un valore dell'argomento pari a 0, come si può notare in figura 28. Un vantaggio della funzione sigmoide consiste nel fatto di essere una funzione continua, a differenza della funzione a gradino, per cui è possibile calcolarne la derivata in qualsiasi punto lungo la curva. Storicamente è stata molto utilizzata poiché rappresenta un'interpretazione molto veritiera del cosiddetto firing rate neuronale, ovvero del tasso di innesco dell'attività di un neurone biologico. Dal punto di vista matematico, tale funzione è descrivibile con

$$\sigma = \varphi(\nu) = \frac{1}{1 + e^{-\nu}}$$

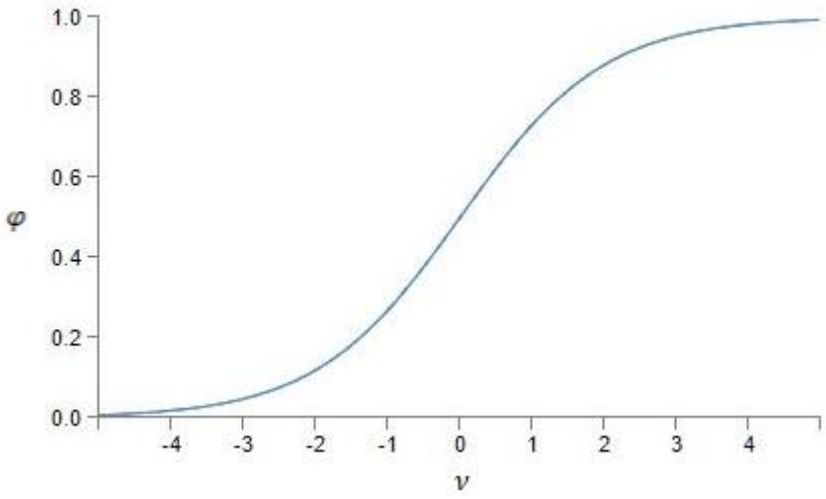

Figura 2811: funzione sigmoide

- Funzione tangente iperbolica (Tanh): tale funzione è molto simile alla precedente, ma si differenzia dalla funzione sigmoide in quanto, come è possibile notare in figura 29, risulta centrata sullo zero e l'argomento della funzione viene, perciò, trasformato in un numero compreso nell'intervallo tra -1 e 1. La tanh deriva da una traslazione della funzione sigmoide

secondo la formula per cui $\tanh(v) = 2\sigma(2v) - 1$. Dal punto di vista matematico, la funzione tangente iperbolica è descrivibile con

$$tanh = \varphi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}}$$

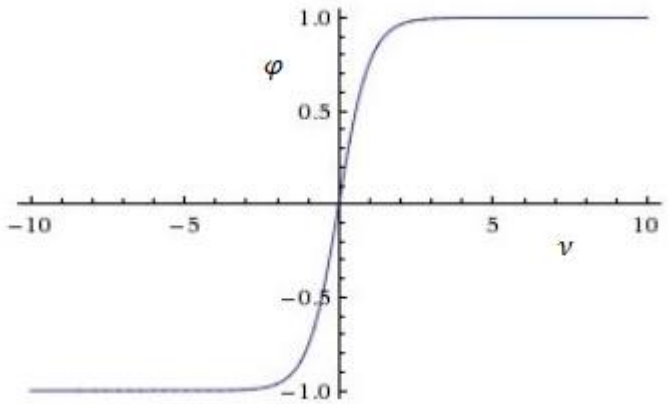

Figura 29: funzione tanh

- Rectified Linear Unit (ReLU): tale funzione è particolarmente utilizzata per la semplicità di calcolo che la caratterizza, in quanto se l'argomento è negativo allora il risultato della funzione sarà 0, altrimenti l'argomento rimarrà invariato, come si può notare in figura 30. Per sua natura, la ReLU filtra i valori negativi in ingresso non prendendoli in considerazione e questo spesso consente di velocizzare sensibilmente la fase di addestramento della rete neurale. Matematicamente può essere espressa tramite la formula

$$ReLU = \varphi(v) = max\{0\,, v\}$$

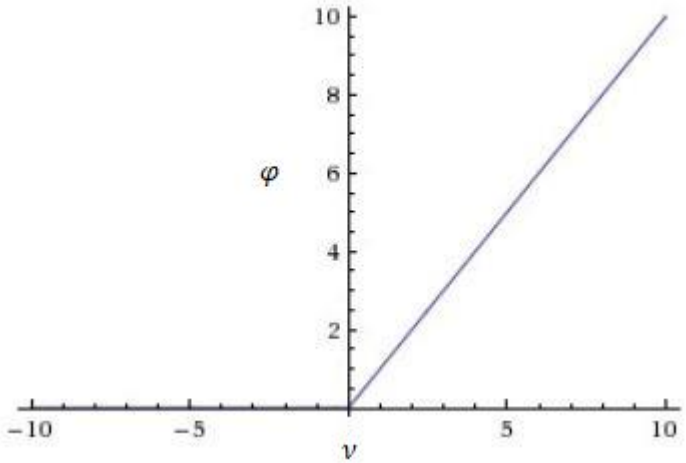

Figura 3012: funzione ReLU

- Leaky ReLU: tale funzione deriva da una modifica della ReLU per eliminare il problema definito dying ReLU. Quest'ultimo si riferisce al fatto che quando un neurone, a cui è applicata una funzione di attivazione ReLU, riceve in input un valore negativo, allora

continuerà ad emettere un output nullo per tutto il funzionamento della rete neurale, non giocando alcun ruolo nella discriminazione dell'input, diventando quindi inattivo e inutile. La funzione Leaky ReLU, come si può osservare in figura 31, consente dunque di limitare l'azione di filtraggio dei valori negativi in input, applicando a questi ultimi un coefficiente α solitamente minore di 1. Dal punto di vista matematico, la funzione Leaky ReLU è riassumibile con

$$Leaky\ ReLU = \varphi(v) = \begin{cases} v, & v \geq 0 \\ \alpha v, & v < 0 \end{cases}$$

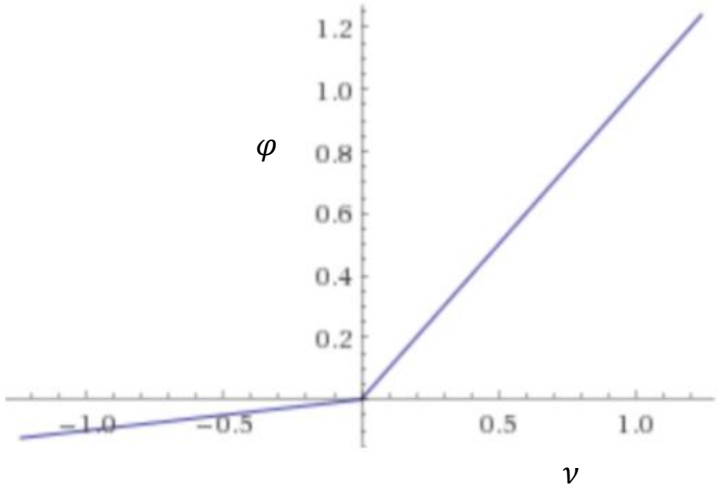

Figura 31: funzione Leaky ReLU

3.3.2 SIMPLE E DEEP NEURAL NETWORK

Le unità computazionali, da cui sono costituite le reti neurali artificiali, si suddividono, in base al layer in cui si trovano, in unità di input, unità nascoste e unità di output.

Le unità di input, che costituiscono il layer iniziale, rappresentano le variabili indipendenti del problema in analisi, che verranno elaborate dalla rete per giungere alla soluzione del problema. La quantità di unità di input è, dunque, definita dal numero di funzionalità intrinseco nel set di dati di input.

Le unità nascoste costituiscono quelli che sono definiti hidden layer, denominazione che deriva dal fatto che i valori di output di tali unità non sono presenti nei dati e non sono perciò visibili. Le unità nascoste hanno l'obiettivo di determinare quali sono quelle informazioni che risultano utili per spiegare le relazioni presenti nei dati osservati, creando una rappresentazione del problema all'interno della rete, ponderando e perfezionando i calcoli elaborati negli strati precedenti tramite algoritmi progressivamente più complessi, per giungere ad un'astrazione sempre più accurata della soluzione del problema. Poiché non esiste alcuna regola per la determinazione del numero di unità elaborative che devono comporre tale strato, solitamente si procede per tentativi sulla base dell'esperienza, tenendo in considerazione che, comunque, esiste una relazione tra la quantità di nodi nascosti e il numero di nodi di input, la quantità di dati utili all'addestramento della rete e la complessità computazionale del problema considerato.

Infine, le unità di output, che costituiscono il layer finale, rappresentano le variabili dipendenti che definiscono la soluzione del problema. Proprio per il fatto che le unità di output svolgono la funzione

di calcolare l'identificazione finale, la classificazione o la previsione di un determinato modello, tali neuroni, a differenza di tutti gli altri che compongono la rete, non presentano una funzione di attivazione, oppure è possibile considerarli come provvisti di una funzione di attivazione lineare identica.
I layer di input e di output sono definiti layer visibili, in quanto contengono variabili che è possibile osservare, differenziandoli, così, dai layer nascosti.
Quando si definisce una rete neurale si fa riferimento ad essa tramite il numero di strati da cui è composta, ma in questa enumerazione non bisogna considerare lo strato di input. Infatti, una rete neurale a singolo strato rappresenta una rete senza strati nascosti, ovvero una rete in cui l'input viene direttamente mappato in output. Perciò, metodi come la regressione logistica ad esempio possono essere trattati come casi speciali di reti neurali a singolo strato.
In base a quanti hidden layer sono presenti nella rete si parla di simple neural network, se ve ne è solamente uno, o di deep neural network, quando ve ne è più di uno, e tale differenza si può osservare in figura 32. Le reti neurali in cui sono presenti più di un hidden layer sono proprio quei modelli che hanno portato all'utilizzo del termine deep learning.

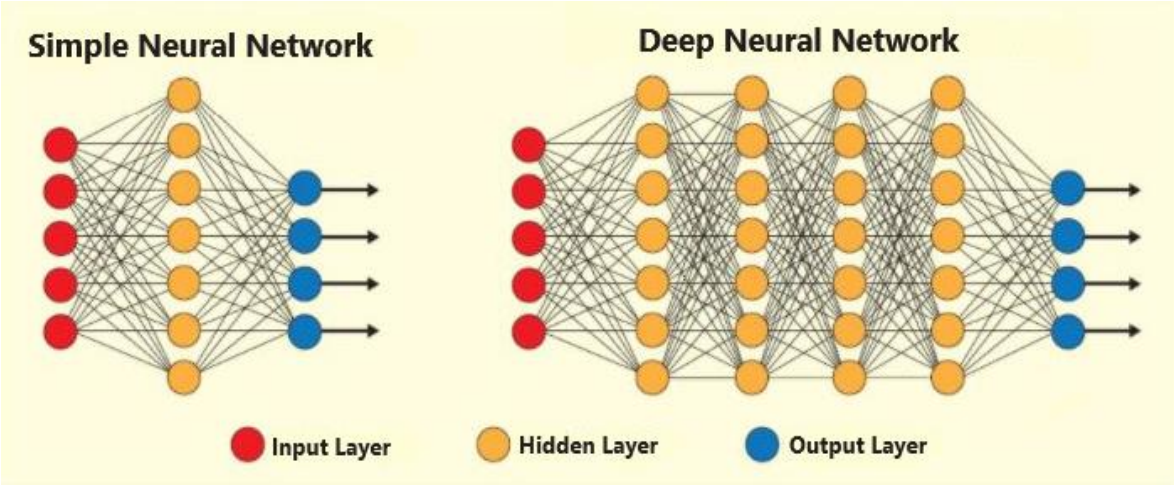

Figura 32: simple NN e deep NN

L'introduzione degli hidden layer ha condotto la comunità dei ricercatori nel campo del deep learning a valutare quanti di questi strati nascosti fossero necessari per l'adempimento di un determinato compito. Nel 1989, il professore di ingegneria al Dartmouth College, George Cybenko, ha elaborato un teorema denominato universal approximation theoreme, all'interno del suo articolo dal titolo "Approximation by superpositions of a sigmoidal function", pubblicato sulla rivista Mathematics of Control, Signals, and Systems, secondo il quale una rete neurale, anche se costituita da un singolo hidden layer, ha la potenza computazionale necessaria per modellare e apprendere il processo di risoluzione per qualunque tipo di problema, anche il più complesso, rispettando comunque determinate condizioni di architettura della rete neurale.
Però, successivamente, con l'avvento di computer sempre più performanti, è stato dimostrato che il numero di unità che compongono una simple neural network aumenta esponenzialmente con l'incremento della complessità dei compiti e tale caratteristica comporterebbe la necessità di una dimensione estremamente elevata per una simple neural network affiché questa possa essere utile alla risoluzione di un determinato problema. Tale complicazione, relativa alla dimensionalità della rete, risulta risolvibile attraverso una disposizione dei neuroni dello strato nascosto su più hidden layer anziché su uno solo, nella maniera che si ritiene più congeniale. In tal modo, si può ottenere

lo stesso risultato, ovvero è possibile sfruttare la medesima potenza rappresentativa e computazionale di una simple neural network di dimensioni gigantesche, riducendo al contempo il numero di unità elaborative e di pesi che il modello deve calcolare, migliorando, così, la flessibilità della rete neurale. Questi risultati sono stati ottenuti grazie ai progressi tecnologici che, superando gli ostacoli rappresentati dalle limitazioni di calcolo, hanno permesso la costituzione di reti neurali con più hidden layer.

L'aumento del numero di hidden layer ha un profondo impatto sui risultati ottenuti dalle deep neural network, in quanto molteplici livelli di astrazione garantiscono alle reti neurali profonde un vantaggio enorme nell'imparare a risolvere complessi problemi di riconoscimento di pattern, proprio perché ad ogni livello intermedio vengono aggiunte informazioni e analisi utili a fornire un output affidabile. Infatti, eseguendo un numero maggiore di calcoli e facendo un numero maggiore di connessioni, i modelli di deep learning permettono di ottenere previsioni che risultano di gran lunga più precise rispetto a quelle dei modelli più semplici. L'effetto sull'accuratezza e sulla migliore generalizzazione dei modelli provocato dalla profondità della rete, a sua volta misurata sulla base del numero di hidden layer presenti, è stata analizzata dall'informatico Ian Goodfellow, insieme ad altri suoi colleghi, i quali nel 2014, durante l'International Conference on Learning Representation, hanno mostrato il grafico rappresentato in figura 33, secondo cui più una rete risulta profonda migliore sarà l'accuratezza della soluzione del problema per cui la rete è stata costituita.

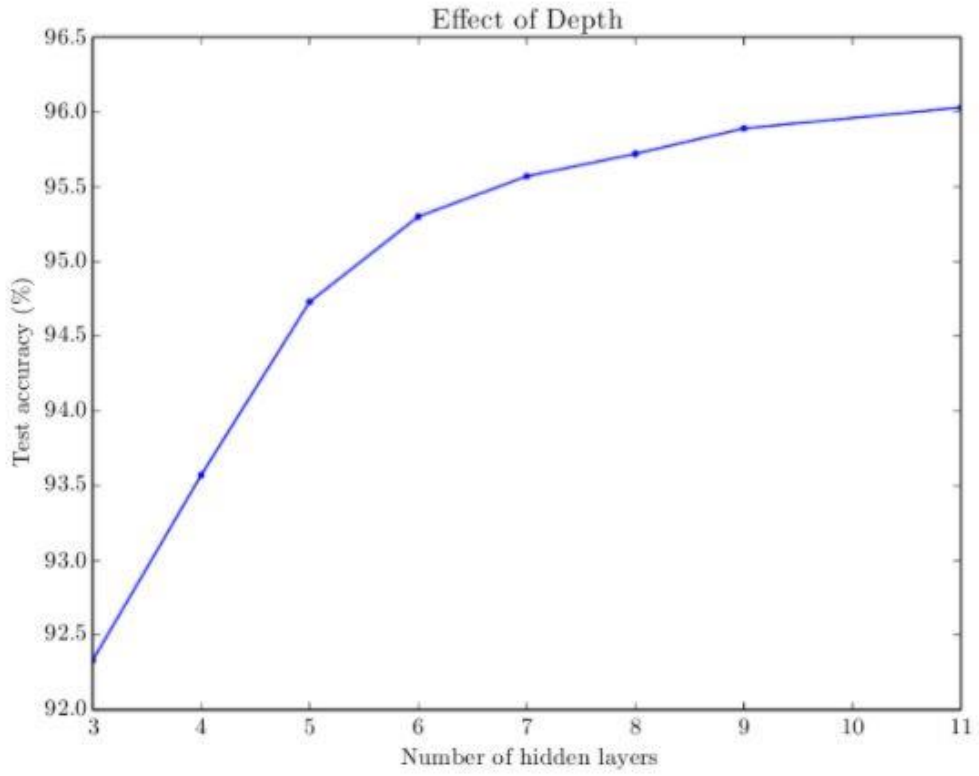

Figura 33: effetto della profondità sull'accuratezza del modello

L'aumento del numero di hidden layer ha permesso ai modelli di deep learning di elaborare una quantità sempre maggiore di dati, dimostrando come questi modelli siano altamente scalabili. La caratteristica della scalabilità è un tratto distintivo delle reti neurali artificiali rispetto agli altri modelli di machine learning. Infatti, i modelli di deep learning migliorano le proprie prestazioni all'aumentare dei dati, riuscendo ad effettuare correzioni estremamente fini sulla base di nuovi dati di input, mentre le altre applicazioni di machine learning, una volta raggiunto un determinato livello

di performance, non sono in grado di ottenere prestazioni superiori, neanche aggiungendo ulteriori dati di addestramento. Inoltre, la necessità di avere una grande quantità di dati, per sfruttare nel migliore dei modi i vantaggi di un modello di deep learning, ha l'effetto di creare un periodo di miglioramento più esteso nel tempo per i modelli di apprendimento profondo, che hanno, perciò, maggiori e migliori possibilità di addestramento.
Nel prosieguo di questo lavoro di tesi, l'attenzione verrà focalizzata sulle deep neural network.

3.3.3 ADDESTRAMENTO DI UNA RETE NEURALE

La fase di addestramento di una deep neural network rappresenta il passaggio fondamentale nella realizzazione di una rete neurale efficiente e si verifica tramite due procedimenti principali di propagazione delle informazioni attraverso la rete. Successivamente, nella trattazione dell'argomento, saranno considerate solamente le reti neurali supervisionate, in quanto le applicazioni finanziarie si basano principalmente su tale tipologia.
Una deep neural network riceve una certa quantità di esempi rappresentativi del problema da affrontare, con le corrispondenti soluzioni. Poiché non è sempre possibile identificare con precisione i fattori che incidono su un determinato fenomeno, può risultare conveniente abbondare nella somministrazione delle variabili di input, le quali, durante l'addestramento, saranno differenziate autonomamente dalla rete in quelle maggiormente significative, alle quali verranno assegnati pesi più consistenti, e quelle marginali, a cui, invece, verranno assegnati pesi ridotti. Una volta ricevute le variabili di input, avviene il procedimento denominato forward propagation, per cui il valore di output della funzione di attivazione di ogni neurone, calcolato ponderando gli input in base ai pesi sinaptici e aggiungendovi il bias, si propaga nei layer successivi fino ad arrivare all'output layer. La rete supervisionata, a questo punto, esamina gli esempi proposti e produce delle ipotesi di soluzione. Tali soluzioni teoriche vengono confrontate con quelle effettive, misurandone lo scostamento tramite una funzione di costo (loss function). Poiché l'obiettivo della neural network consiste nel definire un modello che sia il più efficiente possibile e che generalizzi bene la soluzione, durante la fase di addestramento è necessario un processo di minimizzazione della funzione di costo. La funzione di costo risulta minimizzata quando la soluzione teorica è il più simile possibile a quella effettiva. Il procedimento di minimizzazione dell'errore, commesso nella previsione generata dalla rete, avviene grazie all'utilizzo di due differenti algoritmi: il primo consiste nella back-propagation, a cui ci si riferisce spesso con il termine più conciso backprop, grazie al quale le informazioni relative all'errore misurato dalla funzione di costo vengono propagate nella rete nel verso contrario a quello percorso precedentemente durante la forward propagation, ovvero a partire dai neuroni dello strato di output, passando per quelli degli hidden layer, fino allo strato di input; il secondo, definito gradient descent (discesa del gradiente), permette di calcolare il gradiente della funzione di costo rispetto ai parametri che la compongono, tramite una valutazione numerica che risulta relativamente poco costosa in termini di potenza computazionale grazie all'utilizzo del metodo definito chain rule, e di aggiornare i parametri della funzione di costo, ovvero i pesi sinaptici, nella direzione negativa del gradiente in modo tale da minimizzare la differenza tra i risultati teorici del modello e quelli effettivi.
La chain rule è necessaria per rendere il processo di minimizzazione dell'errore meno oneroso dal punto di vista computazionale, in quanto permette di scomporre il gradiente della funzione di costo in tutte le derivate parziali che la compongono, le quali a loro volta vengono sottoposte allo stesso

procedimento di scomposizione, il quale si reitera un numero elevato di volte per arrivare ad analizzare l'effetto che ogni parametro, di cui è costituita la rete e che influenza la funzione di costo, ha sulla funzione di costo stessa.
L'aggiornamento dei pesi nella direzione negativa del gradiente implica il fatto che il nuovo peso, calcolato ogni volta che avviene il processo della back-propagation, risulti pari alla differenza tra il peso precedentemente utilizzato e la derivata parziale della funzione di costo rispetto al parametro in considerazione, la quale viene moltiplicata per uno scalare definito come tasso di apprendimento (learning rate). Considerando $w_1$quale parametro da aggiornare, la formula definita dall'algoritmo gradient descent risulta dunque

$$w_1^{new} = w_1^{old} - \alpha \nabla_{w_1^{old},L}$$

dove α rappresenta il tasso di rendimento, e ∇ è l'operatore nabla che rappresenta il gradiente del parametro w_1 rispetto alla funzione di costo L.
Poiché il gradiente rappresenta la pendenza di una funzione, in questo caso della funzione di costo, il gradient descent, come si può vedere in figura 34, comporta una "discesa" lungo la pendenza di tale funzione, tramite l'aggiornamento progressivo dei parametri, fino a raggiungere il minimo globale della funzione.

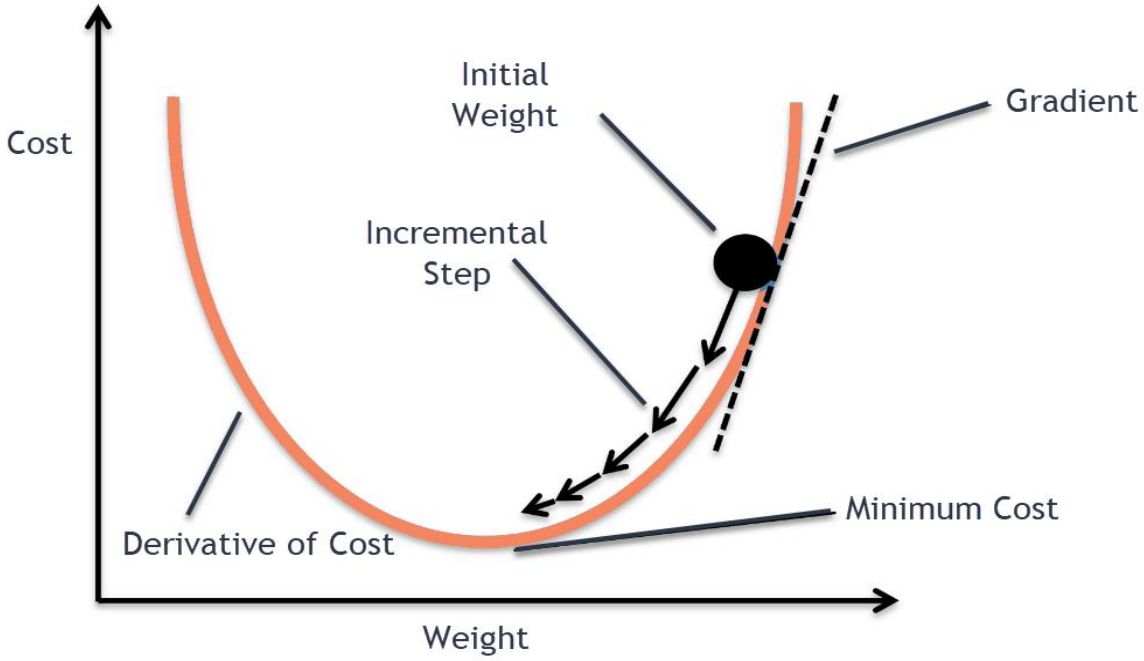

Figura 34: raffigurazione del procedimento del gradient descent algorithm

Il learning rate rappresenta lo spostamento consentito ai pesi delle connessioni nella direzione negativa del gradiente, ovvero definisce la rapidità con cui vengono aggiornati i valori dei pesi e di conseguenza definisce il tempo necessario alla rete per completare il processo di apprendimento. La scelta del tasso di apprendimento deve essere, però, ponderata attentamente, in quanto valori troppo ridotti o troppo elevati di learning rate conducono ad un apprendimento significativamente lento: solitamente, il learning rate viene impostato ad un valore pari a 0.001.

Una volta modificati i pesi, si ripete il susseguirsi di forward e back-propagation, e, ad ogni iterazione, viene replicato il gradient descent algorithm. Ogni ciclo completo di confronto delle soluzioni reali di tutti gli esempi con quelle generate dalla rete prende il nome di epoca.
L'addestramento dei modelli di deep learning risulta fortemente condizionato dai valori iniziali che assumono i pesi sinaptici, in quanto tali valori possono rendere possibile o meno la convergenza dell'algoritmo di gradient descent verso il minimo globale della funzione di costo. Infatti, esistono valori iniziali che risultano tanto instabili da generare difficoltà computazionali così impegnative che l'algoritmo non risulta in grado di risolverle e fallisce nel trovare la soluzione del problema in analisi. Inoltre, i valori iniziali dei pesi, una volta ottenuta la convergenza dell'algoritmo, possono determinare la velocità con cui converge l'apprendimento e possono anche influenzare la generalizzazione del modello.
I pesi iniziali, che vanno assegnati agli input dei neuroni, costituiscono uno dei numerosi iper-parametri, ovvero quei parametri per i quali l'impostazione dei valori non viene appresa automaticamente dalla rete neurale e che, dunque, devono essere definiti inizialmente in fase di designazione dell'architettura della rete stessa. Se due unità, con la stessa funzione di attivazione, sono collegate agli stessi input, esse devono necessariamente avere pesi iniziali diversi, in quanto se avessero parametri identici allora il modello aggiornerebbe costantemente queste unità nello stesso modo. Per questa proprietà, che implica la rottura della simmetria tra gli aggiornamenti di unità differenti, risulta dunque necessaria un'inizializzazione dei pesi del modello a valori casuali tratti da una distribuzione gaussiana o uniforme. In alternativa, per soddisfare tale proprietà, si potrebbe adottare una funzione di attivazione diversa per ogni neurone, ma tale soluzione si rivela molto costosa dal punto di vista computazionale, e dunque viene preferita l'inizializzazione casuale dei pesi, mentre i bias vengono impostati a valori costanti scelti euristicamente.
Le due problematiche più comuni che si possono riscontrare quando non avviene la convergenza dell'algoritmo sono denominate vanishing ed exploding gradient. La prima si verifica quando, a causa di pesi inizializzati a valori estremamente ridotti, il gradiente tende a ridursi man mano che l'algoritmo di back-propagation propaga l'informazione portata dalla funzione di costo negli strati precedenti. Ciò comporta un apprendimento sempre più lento per le unità dei layer iniziali, giungendo nel peggiore dei casi anche a circostanze in cui i pesi non vengono aggiornati e il processo di apprendimento non si verifica. La seconda problematica deriva dalla causa opposta, ovvero da un'inizializzazione dei pesi a valori estremamente elevati, che comporta un aumento sempre maggiore del valore del gradiente, da cui la definizione di exploding gradient, ma determina la medesima conseguenza, ovvero un procedimento di apprendimento che avviene in maniera lenta fino a non riuscire a raggiungere il minimo della funzione di costo. Tali problemi di instabilità del gradiente possono essere risolti attraverso una corretta inizializzazione dei pesi.
L'addestramento termina dopo un certo numero di epoche prefissato dallo sperimentatore oppure dopo che il margine di scostamento tra soluzioni calcolate e soluzioni effettive si sia progressivamente ridotto a un livello predefinito.
Ultimata la fase di addestramento, si passa a quella di verifica con la somministrazione di alcuni esempi del problema, diversi da quelli già utilizzati, le cui soluzioni, questa volta, non vengono comunicate. La rete, sulla base dei pesi elaborati durante l'addestramento, fornisce le proprie soluzioni, la cui validità è strettamente dipendente dal livello di efficienza raggiunto.
Se l'esito è soddisfacente, la rete è pronta per essere utilizzata nel settore di applicazione per il quale è stata creata, al fine di effettuare delle previsioni su valori futuri.

3.3.4 TIPOLOGIE DI RETI NEURALI

Esistono numerose tipologie di deep neural network, ma realizzare una lista esaustiva risulta praticamente impossibile, a causa del fatto che vengono continuamente costituite nuove architetture di reti neurali, ed inoltre è al di fuori dello scopo del presente lavoro di tesi. Di seguito, viene presentata una breve panoramica su alcune delle reti neurali più conosciute ed utilizzate:

- Feed Forward Neural Networks (FF o FFNN): le FFNN sono reti neurali molto semplici, costituite da layer le cui unità non sono mai connesse con altre unità dello stesso strato, ma in cui, generalmente, due layer adiacenti risultano completamente connessi (fully-connected), ovvero ogni neurone di uno strato si collega ad ogni neurone dello strato successivo. L'uso di queste semplici reti neurali è molto limitato, ma spesso vengono combinate con altre reti per formarne di più complesse, al fine di risultare utili per una più vasta gamma di applicazioni.
- Auto Encoder (AE): gli AE derivano dalle FFNN e rappresentano un uso diverso di queste ultime, piuttosto che una diversa architettura di rete neurale. Infatti gli AE hanno l'obiettivo di comprimere le informazioni, codificandole a livello degli hidden layer. Le reti AE assumono la forma di una clessidra, in quanto gli strati nascosti, dove viene compressa l'informazione, risultano dimensionalmente più ridotti rispetto agli strati di input e di output. La porzione della rete composta dagli strati che vanno dall'input layer fino all'hidden layer centrale viene chiamata parte di codifica, il layer centrale rappresenta il codice e la restante parte la decodifica. Tale rete neurale gode della proprietà di simmetria, data dalla forma della rete stessa, che può essere applicata anche ai pesi, facendo sì che i pesi di codifica siano uguali ai pesi di decodifica.
- Sparse Auto Encoder (SAE): i SAE possono essere identificati come l'opposto degli AE, in quanto rappresentano le informazioni in uno spazio maggiore, anziché codificare e comprimere le informazioni in un numero ridotto di nodi. Dunque, nelle reti SAE i layer centrali risultano molto più grandi rispetto ai layer di input e di output. Tali reti neurali possono essere utilizzate per estrarre caratteristiche particolarmente fini dal dataset iniziale. L'addestramento dei SAE è diverso rispetto a quello degli AE, in quanto gli errori vengono propagati nella rete tramite la back-propagation, ma ad essi viene aggiunto un elemento definito sparsity driver, che rappresenta un filtro, grazie al quale solo determinati errori risultano rilevanti ai fini dell'addestramento, mentre altri assumono un valore nullo e non vengono considerati durante il ciclo di back-propagation che si sta attuando in quel determinato momento.
- Recurrent Neural Networks (RNN): le RNN permettono di creare dei feed-back e dei cicli temporali, grazie al fatto che l'output di un hidden layer in un determinato istante di tempo può essere usato come ulteriore input dallo stesso hidden layer per calcolarne lo stato nell'istante di tempo successivo. Le profondità di questa tipologia di reti neurali si riscontra lungo l'asse temporale e tra i vari hidden layer presenti nel recurrent layer. Le RNN possono essere utilizzate in molti campi, in quanto vi sono molti dati i cui input e output necessitano di essere ordinati secondo una sequenza temporale.
- Convolutional Neural Networks (CNN): le CNN, definite anche Deep Convolutional Networks (DCNN), sono reti neurali che rappresentano lo stato dell'arte per la classificazione delle immagini, il riconoscimento e l'identificazione di oggetti, volti, individui e molti altri aspetti tipici dell'elaborazione visiva. Le CNN ricevono in input le immagini,

intese come matrici di numeri, e trasmettono l'informazione attraverso più layer, al fine di eseguire l'estrazione e la classificazione delle caratteristiche più importanti presenti nelle matrici di input. Il principale punto di forza delle CNN è il ridotto numero di pesi necessari a definire l'architettura della rete, in quanto, a differenza delle fully-connected network, ogni neurone risulta collegato solo ad alcuni neuroni degli strati successivi, anziché a ciascuno di essi. Inoltre, i pesi sono condivisi cosicché neuroni con la stessa caratteristica in diverse aree dell'immagine inneschino la stessa risposta. In altre parole, la condivisione dei pesi permette alla rete di riconoscere una geometria, appresa in una determinata sezione dell'immagine, anche in altre parti dell'immagine. Le CNN sono caratterizzate dalla presenza di due specifici tipi di layer. Uno è noto come layer convoluzionale, che prende il nome dall'operazione di convoluzione che viene effettuata tra la matrice di input e il filtro e che consiste nello scorrimento di quest'ultimo sulla matrice in ingresso. Tale operazione produce come risultato una feature map, ovvero una mappa delle caratteristiche specifiche che i nodi della rete tentano di classificare. L'esecuzione di numerose convoluzioni sull'input genera diverse feature map, la cui sovrapposizione permette di ottenere l'output finale del layer convoluzionale, con un livello di astrazione proporzionale alla profondità del layer stesso. L'altro tipo di layer è denominato pooling layer, il quale compie un'operazione di partizione della matrice di input in un insieme di set non sovrapposti per trovare il valore massimo di input in ciascuna regione, risultando così utile a diminuire la quantità di parametri della rete, al fine di ridurre la dimensione spaziale della rete.

- Deconvoutional Networks (DN): le DN, chiamate anche Inverse Graphics Networks (IGN), effettuano procedimenti opposti rispetto alle CNN a livello di hidden layer, svolgendo principalmente operazioni di interpolazione ed estrapolazione, e vengono addestrate a produrre solamente un determinato output già presente tra i diversi dati di input, classificati in vettori binari.
- Deep Convolutional Inverse Graphics Networks (DCIGN): le DCIGN rappresentano un particolare tipo di AE con CNN e DN come codificatore e decodificatore rispettivamente. Queste reti tentano di modellare le caratteristiche presenti nella parte di codifica della rete come probabilità e possono imparare ad effettuare trasformazioni complesse sulle immagini, come ad esempio modificare la fonte di luce o la rotazione di un oggetto 3D.
- Generative Adversarial Networks (GAN): le GAN consistono in una rete formata da due reti neurali, spesso una FFNN e una CNN, di cui una, denominata rete generativa, ha il compito di creare contenuti e l'altra, denominata rete discriminante, riceve in input i dati di addestramento e i dati prodotti della rete generativa. Il compito della rete discriminante consiste nel riuscire a distinguere la fonte dei dati, differenziando nel migliore dei modi quali dati sono reali e quali invece derivano dalla rete generativa, la quale utilizza la qualità della discriminazione come errore per migliorarsi nella generazione di contenuti il più possibile simili a quelli reali. Tale meccanismo produce una competizione tra le due parti della rete, per cui la rete discriminante cerca di migliorare nel distinguere i dati e la rete generativa impara a diventare meno prevedibile per la discriminante. Le GAN risultano molto difficili da addestrare, poiché, oltre all'addestramento di due reti, è necessario anche bilanciare le dinamiche presenti tra esse. Infatti, se uno dei due procedimenti di generazione e di previsione diventa di gran lunga migliore rispetto all'altro, allora la rete convergerà con estrema difficoltà a causa di una divergenza intrinseca.

- Deep Residual Networks (DRN): le DRN sono FFNN molto profonde, ovvero con numerosi hidden layer, e con connessioni che oltre a trasferire l'output di un layer al layer successivo, passano l'input di quel determinato layer ad uno strato successivo distante da 2 a 5 strati da quello da cui parte l'input. Con tale meccanismo, i vari layer della rete hanno come obiettivo quello di apprendere come mappare un input in un output, a cui si devono aggiungere come ulteriore output gli stessi dati ricevuti in input così come sono, senza che vengano elaborati. È stato dimostrato che queste reti sono molto efficaci nell'apprendere modelli contenenti un elevato numero di hidden layers.

3.3.5 RETI NEURALI RICORRENTI

Il presente paragrafo è dedicato ad una trattazione più esaustiva delle caratteristiche e delle funzionalità di una rete neurale ricorrente, in quanto essa sarà alla base del modello che verrà sviluppato ed illustrato nel prossimo capitolo.
Le RNN sono una tipologia di reti neurali specializzate per l'elaborazione di situazioni in cui gli input e gli output del modello hanno diverse dimensioni e quando i componenti degli input e degli output sono ordinati secondo una sequenza temporale, anche di lunghezza variabile. Infatti, mentre le FFNN e tutte quelle reti in cui non sono presenti unità ricorrenti sono limitate alla sola possibilità di mappare un input in un output, per quanto riguarda le RNN non è presente questo vincolo. Dunque, le RNN possono avere input e output di varia lunghezza e ciò le rende particolarmente adatte all'utilizzo per diversi casi d'uso, le cui circostanze si evolvono con il passare del tempo e si rivela necessario tenere in considerazione gli stadi passati per una più accurata modellazione, tra cui ad esempio la classificazione dei sentimenti, la traduzione automatica dei testi, l'applicazione di didascalie alle immagini, il riconoscimento vocale e la previsione dei prezzi azionari.
Proprio per tale caratteristica che le contraddistingue, le RNN vengono classificate in:

- RNN uno a uno: reti in cui sia l'input che l'output sono vettori di lunghezza fissa, come si può vedere in figura 35.

Figura 35: RNN uno a uno

- RNN uno a molti: reti in cui l'input è un vettore di lunghezza fissa, mentre l'output può essere di lunghezza variabile, come si può osservare in figura 36, utilizzabile ad esempio per l'applicazione di didascalie alle immagini.

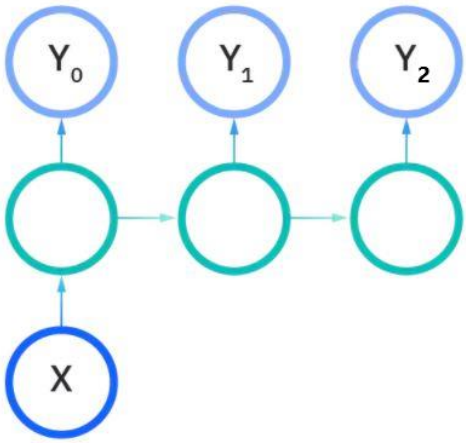

Figura 36: RNN uno a molti

- RNN molti a uno: reti in cui l'input è un vettore di lunghezza variabile, mentre l'output ha una lunghezza fissa, come si può notare in figura 37, utilizzabili ad esempio per la previsione dei prezzi azionari.

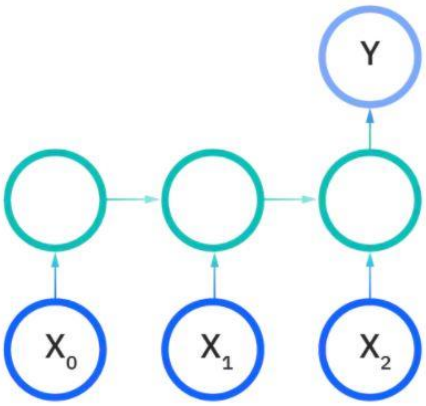

Figura 37: RNN molti a uno

- RNN molti a molti: reti in cui entrambi i vettori di input e di output sono di lunghezza variabile. Vi possono essere due architetture diverse per le RNN molti a molti, in quanto una struttura può avere input e output di lunghezza variabile ma di dimensioni uguali tra loro, come si può vedere nella parte sinistra della figura 38, oppure sequenze di input e output di lunghezza variabile ma con dimensioni diverse tra loro, osservabile nella parte destra della medesima figura.

L'applicazione di un determinato modello a sequenze di dati di lunghezza variabile è resa possibile da una caratteristica peculiare delle RNN, che consiste nel principio di condivisione dei parametri tra tutti gli hidden layer della rete. Mentre gli strati nascosti delle reti feed-forward hanno ognuno i propri pesi e bias, dando così origine ad un comportamento indipendente da parte di ognuno di essi, le RNN condividono tali parametri, permettendo di combinare insieme i vari hidden layer in un unico recurrent layer. Dunque, le RNN convertono le attivazioni indipendenti dei neuroni della rete in attivazioni dipendenti l'una dall'altra, e tale caratteristica permette al modello di ridurre la complessità intrinseca nella natura delle reti neurali, costituita dall'elevato numero di parametri da apprendere e dalla necessità di memorizzare gli output di ogni layer da assegnare come input al layer successivo. La condivisione dei parametri risulta particolarmente importante quando vi sono

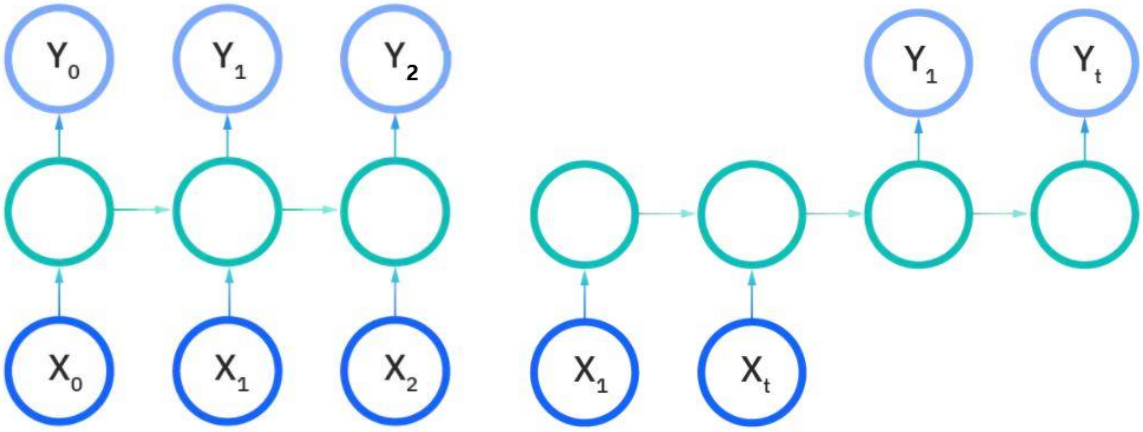

Figura 38: RNN molti a molti

determinate informazioni che possono determinare risposte omogenee in diverse sezioni della sequenza che viene elaborata dalla rete neurale. Se fossero presenti parametri diversi per ogni valore dell'indice temporale, allora non sarebbe possibile generalizzare il modello a sequenze di lunghezze diverse da quelle adottate per l'addestramento della rete, né condividere la forza statistica tra posizioni differenti nel tempo.
La condivisione dei parametri può essere più facilmente intuibile tramite la raffigurazione di una generica RNN molti a molti, osservabile in figura 39, dove la sezione a destra (unfolded RNN) mostra la decompressione della rete presente nella parte a sinistra. Il vettore di input è rappresentato da x, mentre h_t è lo stato dell'hidden layer al tempo t, o è l'output prodotto che va confrontato con quello effettivo y, per generare la funzione di costo L. La particolarità delle RNN consiste proprio nella formula ricorsiva degli stati temporali degli hidden layer, che permette la condivisione dei pesi e che può essere espressa tramite la seguente rappresentazione matematica

$$h_t = \varphi(h_{t-1}, x_t, b)$$

Poiché i dati in ingresso ricevuti da un determinato hidden layer al tempo t sono costituiti dall'input al tempo t, dal bias e dall'output prodotto dallo stesso hidden layer nell'istante di tempo precedente, si viene a creare un ciclo di dipendenza temporale all'interno della rete che porta alla raffigurazione appena illustrata.
La regolazione dei pesi avviene, come nelle altre reti, attraverso i processi di back-propagation e gradient descent, anche se, più nello specifico, le RNN adoperano l'algoritmo di backpropagation through time (BPTT), che risulta leggermente diverso da quello di back-propagation tradizionale, in quanto è specifico per i dati per i quali esiste una sequenza temporale. I principi del BPTT sono gli stessi della back-propagation tradizionale, in cui il modello viene addestrato tramite il calcolo dell'errore e la sua propagazione dallo strato di output verso quello di input, ma la differenza consiste nel fatto che BPTT somma gli errori ad ogni passo temporale, mentre le reti feed-forward non hanno bisogno di sommare gli errori in quanto non avviene la condivisione dei parametri.

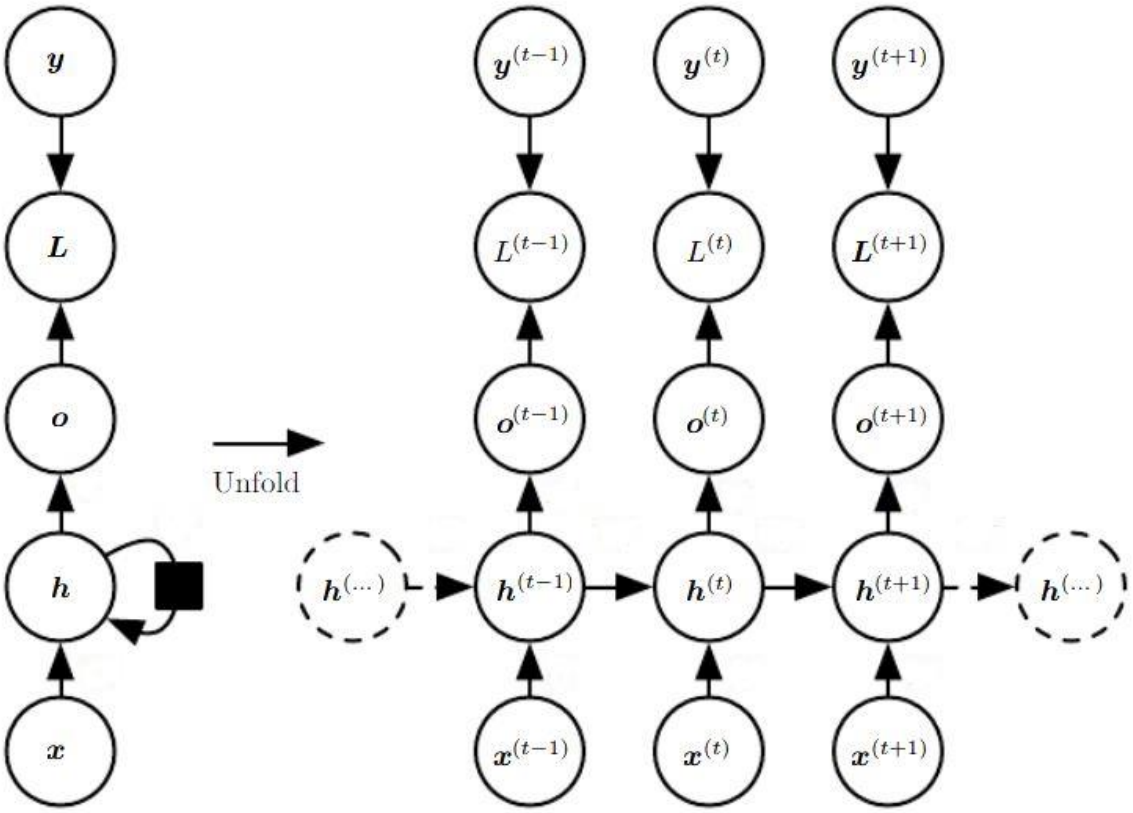

Figura 39: unfolded RNN

Grazie alla struttura ricorsiva degli hidden layer delle RNN e al fatto che il loro obiettivo consista nell'identificare le dipendenze temporali di breve ma soprattutto di lungo termine presenti nei dati di input, queste reti possono essere molto profonde.

La problematica maggiore che si riscontra nell'utilizzo delle RNN è il vanishing o l'exploding gradient problem, già illustrato nel paragrafo 3.3.3. Oltre che a causa di un'errata inizializzazione dei pesi, tale problematica si può verificare anche nei casi in cui la rete neurale sia troppo profonda, in quanto il gradiente, se deve essere propagato indietro fino al layer iniziale passando attraverso numerosi hidden layer, può comportare una convergenza dei pesi a valori nulli o una divergenza dei pesi a valori molto elevati. Poiché, come precedentemente affermato, le RNN possono essere molto profonde e in alcuni modelli i layer devono attraversare numerosi intervalli di tempo, questo problema si rivela ancora più rilevante per le RNN che per le altre tipologie di reti neurali. Nel caso si verifichi il vanishing gradient problem, non avviene l'aggiornamento dei pesi, l'informazione portata dagli input si rivela inutile e viene dunque trascurata la dipendenza temporale. Nel caso si verifichi l'exploding gradient problem, il valore dei pesi tende all'infinto, il gradiente risulta caratterizzato da valori indefiniti e non vi è dunque memoria passata.

La soluzione più comunemente diffusa per l'exploding gradient problem consiste nel modificare la derivata dell'errore prima di propagarla all'indietro attraverso la rete e usarla solo in un successivo momento per aggiornare i pesi. Ridimensionando la derivata dell'errore, anche gli aggiornamenti dei pesi saranno ridimensionati, diminuendo drasticamente la probabilità che essi tendano nel tempo a valori eccessivamente elevati. Esistono due metodi principali per aggiornare la derivata dell'errore, denominati gradient scaling e gradient clipping. Il gradient scaling implica la normalizzazione del vettore del gradiente di errore in modo che la norma del vettore sia uguale a

un valore definito, mentre il gradient clipping implica la forzatura dei valori del gradiente ad un valore minimo o massimo specifico se il gradiente supera un determinato intervallo previsto.
Per quanto riguarda invece il vanishing gradient, la ricerca di una soluzione per tale problematica ha determinato la nascita di due differenti modelli di rete neurale, che si presentano come tipologie particolari di RNN, e sono denominati Long Short Term Memory (LSTM) e Gated Recurrent Unit (GRU).

3.3.5.1 LSTM E GRU

Il modello LSTM è stato proposto da due informatici tedeschi, Sepp Hochreiter e Jürgen Schmidhuber, i quali diffusero le loro ricerche in un articolo intitolato "Long Short-Term Memory", pubblicato sulla rivista Neural Computation nel 1997. Il nome si riferisce all'idea secondo la quale le attivazioni di una rete corrispondono alla memoria a breve termine, mentre i pesi corrispondono alla memoria a lungo termine. L'architettura LSTM è stata progettata al fine di rendere la rete neurale in grado di ricordare e mantenere le informazioni acquisite per lunghi periodi di tempo fino a quando non si rivelano effettivamente necessarie, in modo tale da risolvere il vanishing gradient problem. Infatti, le RNN, che risentono di tale problema, non sono in grado di prevedere accuratamente lo stato attuale nel caso in cui lo stato precedente, che influenza la previsione attuale, non si trovi nell'immediato passato.
L'unità di base delle LSTM, come si può vedere in figura 40, consiste in un blocco di memoria denominato cella. Ogni hidden layer della rete risulta dunque composto da un elevato numero di celle di memoria. Con le LSTM, l'informazione scorre attraverso un meccanismo noto come stato della cella (cell state). In questo modo, le LSTM possono ricordare o dimenticare selettivamente le informazioni che vengono ritenute più o meno rilevanti. L'informazione in un determinato cell state dipende da tre diversi fattori:

- Stato precedente della cella: rappresenta l'informazione che era presente nella memoria alla fine dell'intervallo di tempo precedente, ed è evidenziato dalla linea verde che attraversa le varie celle in figura 40.
- Stato precedente dell'hidden layer (ht-1): rappresenta l'informazione che deriva dall'output della cella precedente.
- Input corrente (xt): rappresenta l'informazione che viene fornita alla rete nell'intervallo di tempo corrente.

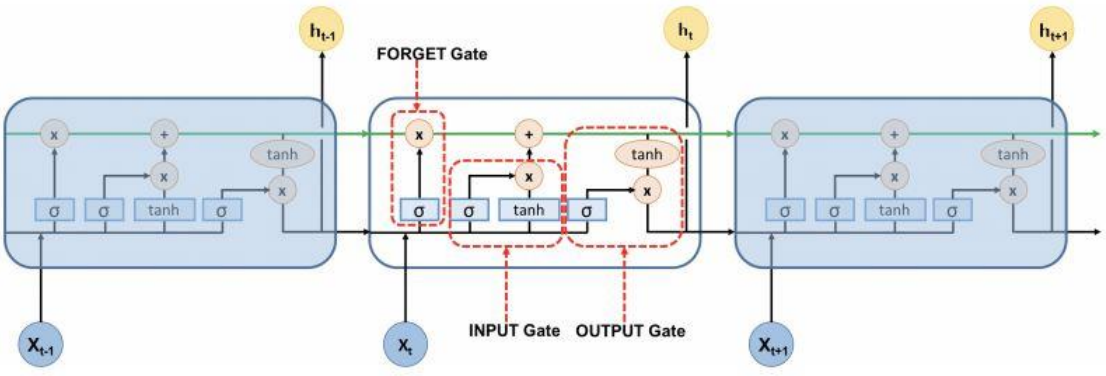

Figura 40: struttura LSTM

Le informazioni che confluiscono nella cella di memoria vengono elaborate attraverso tre meccanismi principali, denominati gate:

- Forget gate: rappresenta l'elemento responsabile della rimozione dallo stato della cella delle informazioni ritenute non importanti. Questo procedimento permette di ottimizzare le prestazioni della rete LSTM. Il forget gate riceve in input ht-1 e xt, li moltiplica per i rispettivi pesi, vi aggiunge il bias e passa il valore ottenuto alla funzione sigmoide, la quale produce un vettore con valori che vanno da 0 a 1, da associare, tramite una moltiplicazione, ai valori dello stato precedente della cella, che rappresentano un ulteriore input per il forget gate. L'emissione da parte del forget gate di un valore nullo per un determinato valore dello stato della cella corrisponde all'eliminazione dallo stato della cella dell'informazione portata da quel valore specifico. Al contrario, l'emissione da parte del forget gate di un valore unitario per un determinato valore dello stato della cella corrisponde alla conservazione da parte dello stato della cella dell'informazione contenuta in esso. Valori intermedi tra 1 e 0 implicano l'assegnazione all'informazione di una rilevanza rispettivamente più o meno marcata.
- Input gate: rappresenta l'elemento che aggiorna lo stato della cella, accertandosi che vengano aggiunte solamente informazioni importanti e non ridondanti. L'input gate riceve in ingresso ht-1 e xt, li moltiplica per i rispettivi pesi, vi aggiunge il bias e passa il valore ottenuto alla funzione sigmoide prima e alla funzione tangente iperbolica in un secondo momento. La funzione sigmoide opera secondo lo stesso principio analizzato in precedenza con il forget gate, secondo il quale produce un vettore di valori che vanno da 0 a 1 e che agisce come un filtro. La funzione tangente iperbolica, invece, produce un vettore contenente tutti i possibili valori, che vanno da -1 a +1, che possono essere aggiunti allo stato della cella. Gli output delle due funzioni vengono moltiplicati tra loro, applicando, dunque, il filtro, definito dalla funzione sigmoide, alle possibili informazioni da mantenere, definite dalla funzione tangente iperbolica. Il risultato ottenuto contiene le informazioni che si ritiene necessario mantenere e viene, perciò, aggiunto allo stato della cella tramite un'operazione di addizione.
- Output gate: rappresenta l'elemento che determina lo stato successivo dell'hidden layer, il quale trasporta le informazioni riguardati gli input precedenti alle celle di memoria successive. L'ouput gate riceve come input ht-1, xt, e lo stato precedente della cella modificato dall'elaborazione appena eseguita da parte del forget gate e dell'input gate. Gli input ht-1 e xt vengono moltiplicati per i rispettivi pesi, vi viene associato il bias e il valore ottenuto viene passato alla funzione sigmoide. L'input derivante dello stato precedente della cella, invece, viene passato alla funzione tangente iperbolica. Moltiplicando i valori ottenuti dalla due funzioni si ottiene il nuovo stato della cella, contenente le informazioni che determinano anche il nuovo stato dell'hidden layer. Il nuovo stato di cella e il nuovo stato dell'hidden layer vengono, dunque, trasmessi alla cella relativa all'istante di tempo successivo.

A causa della complicata architettura della LSTM, tale tipologia di rete neurale ricorrente non ha avuto particolare popolarità fino al 2013-2014, quando i ricercatori sono riusciti ad ottenere con tale modello importanti risultati relativi ai problemi di traduzione automatica e di riconoscimento vocale.

Si sono verificati numerosi tentativi volti a semplificare l'architettura LSTM, che hanno portato alla definizione del modello GRU, molto simile alla LSTM ma più semplice nella struttura e più facile da

addestrare. L’architettura GRU è stata proposta nel 2014 dal professore ed informatico Kyunghyun Cho, insieme ad altri studiosi, nel paper intitolato "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". A differenza della LSTM, la GRU ha solo due gate e l’informazione, che nella LSTM viene conservata nello stato della cella, viene ora incorporata nello stato dell’hidden layer. I due gate presenti in questa architettura sono l’update gate, il quale, analogamente all’output gate della LSTM, determina quanta dell’informazione passata deve essere aggiornata e poi trasmessa agli stati successivi, e il reset gate, il quale, rappresentando una combinazione dell’input gate e del forget gate della LSTM, stabilisce quanta dell’informazione passata non considerare per gli stati delle celle successive.

4. EFFICIENZA DEL MERCATO DELLE CRIPTOVALUTE

Dopo aver eseguito, nei precedenti capitoli, un'analisi dell'efficienza del mercato azionario e delle principali caratteristiche del marcato delle criptovalute, tramite una panoramiche utile a comprendere le funzionalità di queste ultime e la loro portata rivoluzionaria, nel seguente capitolo viene analizzata l'efficienza del mercato delle criptovalute attraverso un confronto con quella del mercato azionario.

Per realizzare tale confronto, vengono presi in considerazione due indici azionari, ovvero Standard & Poor's 500 ed EuroStoxx 50, e il bitcoin, tramite la sua quotazione in dollari.

L'indice S&P 500, creato il 4 marzo 1957, è il più importante indice azionario nordamericano e contiene 500 titoli azionari di altrettante società, quotate alle borse di New York, NYSE e Nasdaq, e rappresentative dell'80% circa della capitalizzazione di mercato. Tutti i titoli in questione sono relativi ad aziende statunitensi che soddisfano una serie di requisiti fondamentali, tra cui una capitalizzazione di mercato superiore a 6,1 miliardi di dollari, un flottante pari ad almeno il 50%, un volume di scambi mensili negli ultimi sei mesi non inferiore a 250.000 azioni, ed un valore medio annuale del titolo superiore a 1 dollaro. Le società da inserire nel paniere vengono selezionate attraverso il metodo della capitalizzazione flottante, dove per flottante si intende la parte di capitale sociale che può essere comprata e venduta sul mercato, ovvero titoli non detenuti dagli Stati, da blocchi di controllo o da azionisti legati da patti di sindacato. La capitalizzazione flottante risulta dunque il prodotto tra il numero di titoli flottanti e i prezzi di tali titoli. L'indice ha un valore che viene calcolato tramite una media ponderata dei prezzi delle azioni delle aziende presenti nel paniere, ed è stato utilizzato come benchmark di riferimento per la maggior parte degli studi effettuati relativamente all'efficienza del mercato azionario, come rilevato nel primo capitolo del presente lavoro di tesi.

L'indice EuroStoxx 50, invece, è stato introdotto il 26 febbraio 1998 e rappresenta le 50 maggiori aziende per capitalizzazione in Europa, che provengono da Belgio, Finlandia, Francia, Germania, Irlanda, Italia, Paesi Bassi e Spagna. La sua composizione, come anche quella dell'S&P 500, è soggetta a revisione con periodicità annuale.

Il bitcoin è stato scelto come riferimento del mercato delle criptovalute in quanto, come si può osservare dal parametro della dominance, la quale si assesta attorno al 62% in data 25 febbraio 2021, rappresenta più della metà della capitalizzazione totale del marcato delle criptovalute. La dominance viene calcolata dividendo la capitalizzazione del bitcoin, la quale ha superato la soglia dei 1000 miliardi di dollari il 20 febbraio 2021, per la capitalizzazione totale del mercato delle criptovalute. La prevalenza di bitcoin, che viene sintetizzata dalla dominance, per quanto riguarda l'assorbimento di quote di mercato rispetto alle altre criptovalute, data soprattutto dal vantaggio del first mover, può essere agevolmente compresa grazie alla figura 41. Inoltre, proprio a causa di tale prevalenza del bitcoin, la maggior parte delle altre criptovalute risulta correlata al bitcoin, seguendone perciò l'andamento generale, seppur con numerose discrepanze nel breve periodo. Dunque, l'analisi del comportamento del bitcoin può risultare utile per una comprensione del mercato di riferimento.

Al fine di individuare l'efficienza relativa del mercato delle criptovalute, è stata creata una rete neurale, argomento le cui peculiarità sono state discusse nel capitolo precedente, con il linguaggio di programmazione Python. Tale rete neurale, la cui struttura verrà esplicitata in seguito, è stata applicata a tutti e tre gli asset considerati, mantenendo la stessa architettura.

Figura 41: dominance BTC (fonte Coin360, dati del 25/02/2021)

Nel mondo finanziario, i modelli di reti neurali vengono spesso utilizzati per applicazioni che hanno come obiettivo la classificazione dei titoli, la modellazione del rischio di credito, la definizione di indicatori e, soprattutto, quello che viene definito trading algoritmico. Nel caso del mercato azionario, ad esempio, i trader generano complessi modelli di deep learning e di reti neurali per identificare azioni sottovalutate o sopravvalutate, e per migliorare ed ottimizzare i modelli esistenti per la gestione dei portafogli di titoli.

Poiché, dunque, le reti neurali si prestano perfettamente alla previsione delle serie temporali, il modello creato per il presente lavoro di tesi viene addestrato con lo scopo di prevedere la futura quotazione dei due indici azionari sopra citati e del bitcoin.

Però, poiché secondo l'ipotesi di mercato efficiente formulata da Eugene Fama, come analizzato nel primo capitolo, l'informazione è disponibile a tutti, tutte le ultime notizie sono prontamente incorporate nei prezzi e questi ultimi dimostrano di avere delle variazioni successive che si comportano secondo un processo di random walk, allora non risulta possibile per gli operatori presenti sul mercato elaborare strategie che permettano loro di ottenere guadagni in eccesso rispetto al normale tasso di profitto, e dovrebbe, dunque, essere impossibile stimare in modo sistematico l'andamento dei prezzi dei titoli quotati sui mercati. Se i prezzi, nella loro formazione, seguono dei processi random walk, che hanno la proprietà di essere senza memoria, per cui la conoscenza degli stati passati in cui si è trovato il sistema non è utile per predirne lo stato futuro, allora la conoscenza dei prezzi passati non risulta utile a predire i prezzi futuri e dunque il valore atteso degli excess returns è nullo.

Nel prosieguo del capitolo il concetto di efficienza verrà identificato con la forma debole, per cui l'analisi della formazione dei prezzi futuri viene fatta considerando solo l'informazione portata dalla serie storica dei prezzi passati, che rappresenta il dato pubblico più diffuso e facilmente conoscibile dagli operatori di mercato.

I risultati del semplice modello di rete neurale proposto, ovvero gli errori commessi nella previsione delle quotazioni future, possono dimostrare, perciò, in modo innovativo, quanto il mercato delle criptovalute possa essere efficiente o meno in confronto a quello azionario, in base alla precisione con la quale i risultati del modello rispecchiano le quotazioni effettive.

4.1 MODELLO IN PYTHON

Innanzitutto è necessario definire le librerie utili allo sviluppo e alla scrittura del modello. La diffusione della popolarità delle reti neurali, avvenuta in maniera considerevole a partire dal 2014, è stata supportata e alimentata proprio dalla pubblicazione online di numerose librerie, note nell'ambito della programmazione, come ad esempio TensorFlow, una libreria software open source per il machine learning, originariamente generata da Google, che fornisce moduli sperimentati e ottimizzati, utili nella realizzazione di algoritmi per diversi tipi di compiti percettivi e di comprensione del linguaggio.
Per il modello in questione viene utilizzata Keras, ovvero un'API di deep learning scritta in Python, progettata come un'interfaccia ad un livello di astrazione superiore rispetto ad altre librerie simili, e che supporta come back-end le librerie TensorFlow. Keras è stata sviluppata con l'obiettivo di essere un'interfaccia accessibile ed altamente produttiva, fornendo astrazioni essenziali e blocchi predefiniti ma flessibili per lo sviluppo e la realizzazione di soluzioni ad alta velocità di interazione per problemi di machine learning, con particolare attenzione al moderno campo del deep learning. Tale libreria rende, dunque, più semplici l'addestramento e la convalida dei modelli di reti neurali.
Inoltre, sono necessarie la libreria NumPy, che rappresenta la crasi di Numerical Python, ovvero il pacchetto fondamentale per il calcolo scientifico con Python, che permette di generare e gestire i dati in modo estremamente rapido, e la libreria pandas, che rappresenta la crasi di panel data, ovvero uno strumento open source di analisi e manipolazione dei dati veloce, potente, flessibile e facile da usare, oltre che tutta un'altra serie di librerie per la lettura e per il plottaggio dei dati, come si può vedere dal codice riportato in figura 42.

```
import pandas_datareader as webreader
import math
import numpy as np
import pandas as pd
from datetime import date, timedelta
from pandas.plotting import register_matplotlib_converters
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from sklearn.metrics import mean_squared_error
from keras.models import Sequential
from keras.layers import LSTM, GRU, Dense, Dropout
from keras.callbacks import EarlyStopping
from sklearn.preprocessing import MinMaxScaler
```

Figura 42: librerie Python

4.1.1 DATI

Per la realizzazione del modello sono stati utilizzati i dati storici relativi alle quotazioni di S&P 500, EuroStoxx 50 e bitcoin presi da Yahoo Finance. Al fine di disporre di un campione di dati simile a livello di numerosità per i tre modelli da analizzare, vengono considerate le quotazioni giornaliere di S&P 500 e di Eurostoxx 50 dall'1 gennaio 2010, mentre per il bitcoin sono analizzati i dati a partire dal 17 settembre 2014, in quanto quest'ultimo, come tutte le altre criptovalute e a differenza degli

indici azionari, ha una quotazione giornaliera che comprende anche i giorni di sabato e domenica. L'ultimo giorno preso in considerazione per lo studio in questione è il 24 novembre 2020. In tal modo, per la rete neurale risultano disponibili 2258 dati giornalieri per il bitcoin, 2730 per l'EuroStoxx 50 e 2744 per l'S&P 500. Per l'intervallo di tempo considerato, i prezzi di chiusura giornalieri definiscono i seguenti grafici:

- Bitcoin:

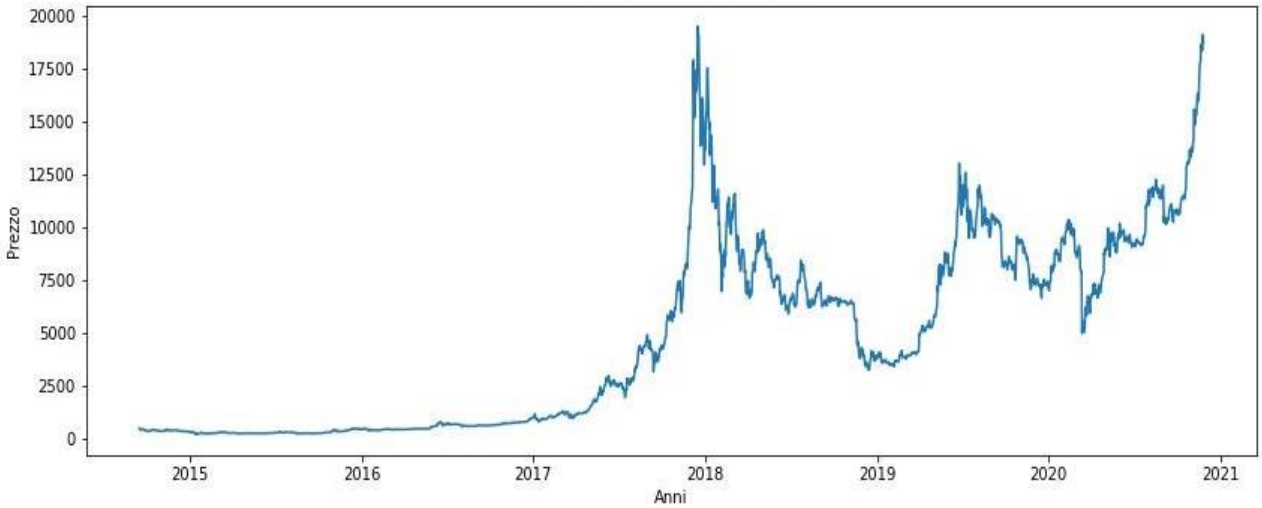

Figura 43: serie storica BTC/USD

- EuroStoxx 50:

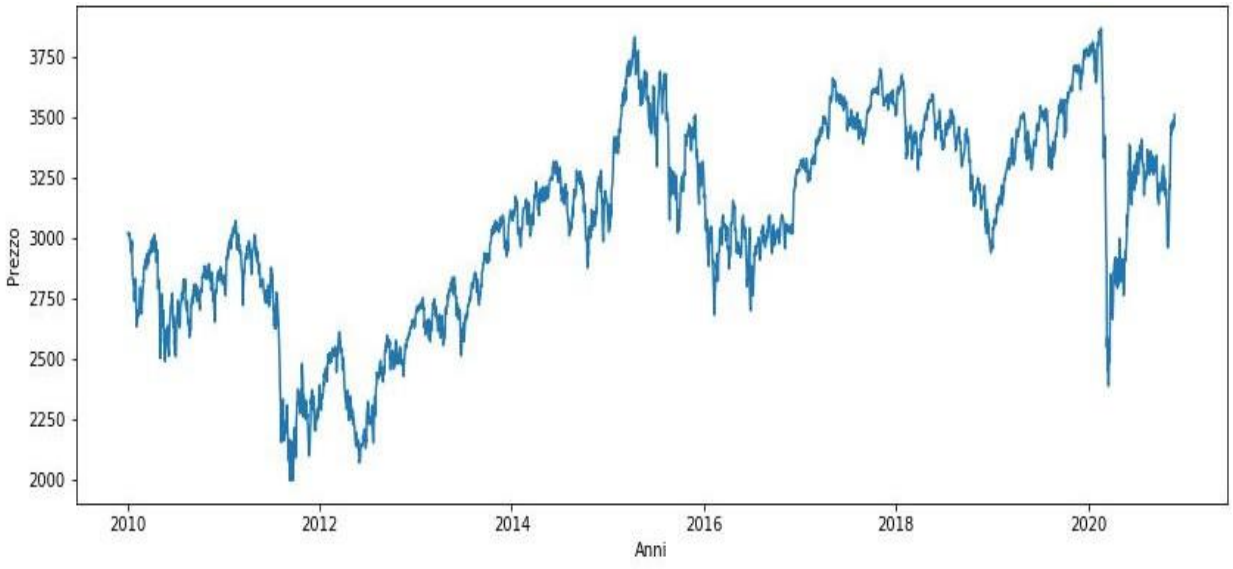

Figura 44: serie storica EuroStoxx 50

- S&P 500:

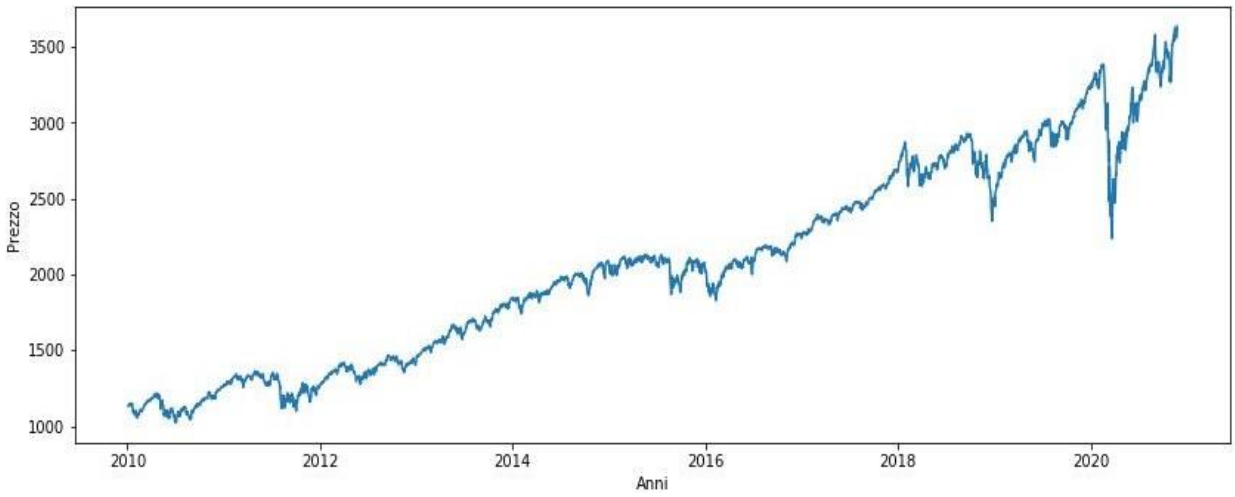

Figura 45: serie storica S&P 500

La lettura dei dati avviene secondo il codice presente in figura 46, nel quale di volta in volta va inserita la dicitura dell'indice su cui si vuole sviluppare il modello e le date relative, in base a quanto riportato sopra.

```
date_start = '2010-01-01' o '2014-09-17'
date_today = '2020-11-24'

stockname = 'EUROSTOXX 50' o 'S&P 500' o 'BTCUSD'
symbol = '^STOXX50E' o '^GSPC' o 'BTC-USD'
df = webreader.DataReader(symbol, start=date_start,
                          end=date_today, data_source="yahoo")
dfs = df.copy()
```

Figura 46: lettura dati da Yahoo Finance

Come si può osservare dal codice in figura 47, per una migliore manipolazione e gestione, si è preferito indicizzare i dati tramite un indice numerico, anziché con l'utilizzo delle date, le quali vengono comunque salvate per un successivo uso nel plottaggio delle previsioni eseguite dal modello.

```
d = dfs.sort_values(by=['Date']).copy()
date_index = d.index
d = d.reset_index(drop=True).copy()
```

Figura 47: indicizzazione dei dati

Le caratteristiche dei prezzi giornalieri che vengono prese in considerazione per la previsione del prezzo futuro, come si può vedere nel codice riportato in figura 48, sono il prezzo massimo, il prezzo minimo, il prezzo di apertura e quello di chiusura, ovvero informazioni a cui ogni operatore di mercato ha facilmente accesso e su cui può effettuare delle analisi semplicistiche sul possibile andamento futuro delle quotazioni. Viene creata una colonna aggiuntiva, definita Prediction, i cui valori inizialmente vengono impostati pari ai valori del prezzo di chiusura per ogni giorno

considerato, e che successivamente saranno aggiornati con i valori previsti dal modello dopo l'addestramento.

```
features = ['High', 'Low', 'Open', 'Close']
data = pd.DataFrame(d)
data_filtered = data[features]
data_filtered_ext = data_filtered.copy()
data_filtered_ext['Prediction'] = data_filtered_ext['Close']
```

Figura 48: definizione del dataframe

La particolare struttura delle reti neurali richiede precise modalità di rappresentazione del problema da risolvere, ovvero tutte le variabili devono trovare rappresentazione in forma numerica per ciascuna unità sia di input che di output. Nel caso di variabili continue, esse devono avere valori compresi tra determinati intervalli numerici. Valori diversi da quelli consentiti devono essere ricondotti alla forma richiesta attraverso un trattamento preliminare di interpolazione. Il ridimensionamento dei dati, definito anche scaling, è un procedimento di pre-elaborazione raccomandato quando si lavora con la maggior parte degli algoritmi di machine learning, in quanto tale approccio rende l'addestramento delle reti neurali più veloce e migliora l'accuratezza delle previsioni. Le due tecniche più popolari per effettuare lo scaling dei dati prima della modellazione consistono nella normalizzazione e nella standardizzazione. La normalizzazione riconduce i dati originali all'interno dell'intervallo di valori compresi tra 0 e 1, e può essere ottenuta per ogni variabile tramite la sottrazione ad essa del valore minimo del set di dati e poi tramite la divisione del valore ottenuto per la differenza tra il massimo e il minimo valore del set di dati. La standardizzazione, invece, ridimensiona ogni variabile di input separatamente sottraendovi la media e dividendo per la deviazione standard, modificando, dunque, la distribuzione dei dati al fine di avere una distribuzione con media nulla e deviazione standard unitaria. Sia la normalizzazione che la standardizzazione possono essere ottenute usando la libreria scikit-learn.
Per il modello in questione, come si può osservare dal codice in figura 49, dopo aver definito il set di dati su cui lavorare e averlo trasformato in un numpy array, viene attuata la normalizzazione tramite l'oggetto MinMaxScaler, sia per quanto riguarda le colonne contenenti le quattro caratteristiche di prezzo prese in considerazione, sia per la colonna aggiuntiva Prediction.

```
np_data_unscaled = np.array(data_filtered)

scaler = MinMaxScaler(feature_range=(0, 1))
np_data = scaler.fit_transform(np_data_unscaled)

scaler_pred = MinMaxScaler(feature_range=(0, 1))
df_Close = pd.DataFrame(data_filtered_ext['Close'])
np_Close_scaled = scaler_pred.fit_transform(df_Close)
```

Figura 49: normalizzazione dei dati

Per permettere alla rete di effettuare l'apprendimento supervisionato, è necessario dividere i dati in due sottoinsiemi, definiti training set e testing o validation set. Per il modello si è deciso di dedicare l'80% del campione dei dati al training set, per l'addestramento della rete neurale, e il restante 20% al testing set, per la verifica della capacità di generalizzazione della funzione di apprendimento. Tale suddivisione in percentuali è una pratica abbastanza comune quando si opera

con l'apprendimento supervisionato. Dunque, per quanto riguarda il bitcoin la rete neurale viene addestrata con 1807 dati giornalieri e testata su 451 giorni, per l'EuroStoxx 50 risultano disponibili 2184 dati per l'addestramento e 546 per la convalida, mentre per l'S&P 500 2196 sono i dati di training e 548 quelli per la verifica del modello.
Per attuare tale suddivisione è necessario, innanzitutto, definire il parametro timesteps, denominato sequence length, che determina il numero di osservazioni precedenti che devono essere considerate dalla rete neurale per effettuare la previsione dell'osservazione futura. Si è deciso di definire una sequence length pari a 100, per cui la rete neurale avrà il compito di prevedere la quotazione futura a seguito dell'osservazione dei prezzi dei cento giorni precedenti.
Il parametro timesteps risulta di fondamentale importanza per le reti neurali, in quanto si preferisce somministrare i dati organizzandoli in intervalli temporali ridotti e tridimensionali denominati batch, dove la prima dimensione rappresenta il numero di campioni, la seconda i timesteps in ogni campione e la terza il numero di feature o caratteristiche della variabile in analisi. Grazie a tale organizzazione, la rete neurale elabora i batch uno per uno e crea una previsione separata per ogni batch. Più nello specifico, la rete neurale elabora i dati dei primi cento giorni compresi nel primo batch per prevedere la quotazione del centounesimo giorno, poi considera il secondo batch in cui sono compresi i dati dal primo al centounesimo giorno per prevedere la quotazione del centoduesimo giorno e così via, in un procedimento che prosegue per il numero di campioni definito dalla prima dimensione del batch, e che può essere compreso dalla rappresentazione in figura 50.

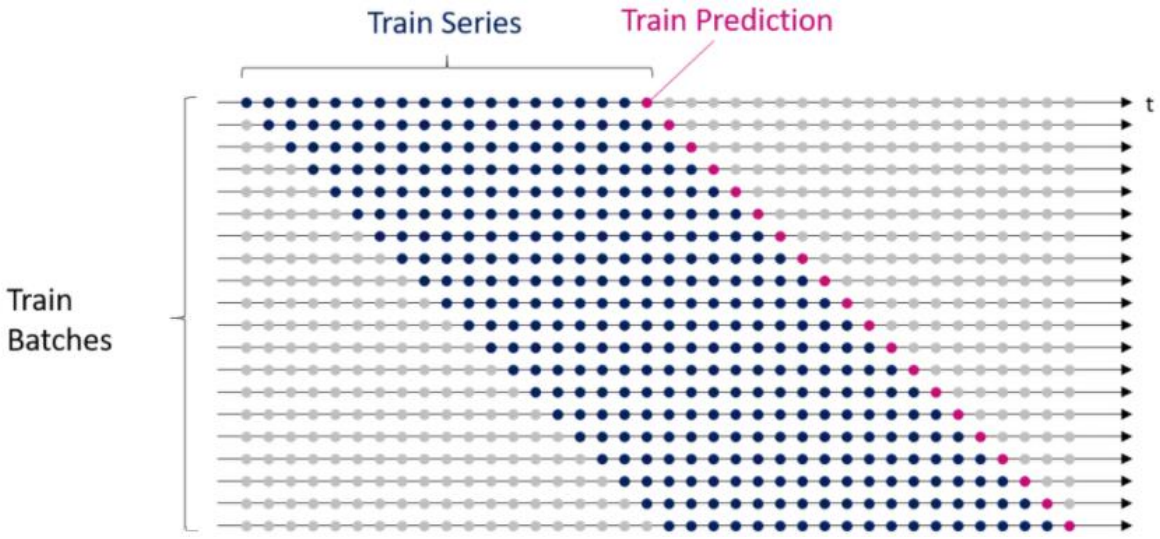

Figura 50: batch e andamento dell'addestramento

Dunque, a loro volta, il training set e il testing set vanno suddivisi in ulteriori raggruppamenti: il training set viene scomposto in x_train e y_train, dove nel primo array sono presenti 1707 campioni per bitcoin, 2084 per EuroStoxx 50 e 2096 per S&P 500, contenenti ciascuno 100 righe e 4 colonne, mentre il secondo è un vettore di una riga e 1707 colonne per bitcoin, 2084 per EuroStoxx 50 e 2096 per S&P 500; il testing set viene invece suddiviso in x_test e y_test, seguendo la stessa logica del training set ma rispettando la dimensionalità di competenza dei dati riservati alla verifica del modello.
La suddivisione dei dati appena esplicitata viene elaborata attraverso il codice in figura 51.

```
sequence_length = 100

training_set_len = math.ceil(np_data.shape[0] * 0.8)
training_set = np_data[0:training_set_len, :]
x_train, y_train = [], []
for i in range(100, training_set_len):
    x_train.append(training_set[i-sequence_length:i,:])
    y_train.append(training_set[i, 0])
x_train, y_train = np.array(x_train), np.array(y_train)

testing_set= np_data[training_set_len - sequence_length:, :]
x_test = []
y_test = np_data_unscaled[training_set_len:, 3:]
testing_set_len = testing_set.shape[0]
for i in range(sequence_length, testing_set_len):
    x_test.append(testing_set[i-sequence_length:i,:])
x_test = np.array(x_test)
```

Figura 51: suddivisione dei dati in training set e testing set

4.1.2 ARCHITETTURA DELLA RETE NEURALE

Poiché risulta necessario operare con dati per cui gli input e gli output hanno diverse dimensioni e per cui i componenti di input e output sono ordinati secondo una sequenza temporale di lunghezza variabile, allora il modello di rete neurale che si presta per tale esigenza consiste nella rete neurale ricorrente. Le reti neurali ricorrenti sono particolarmente utili per l'analisi delle serie temporali e la loro struttura consente l'apprendimento di pattern che si verificano su periodi di tempo diversi.
Al fine di evitare il problema del vanishing grandient e poiché la rete neurale deve essere in grado di ricordare e mantenere le informazioni acquisite per lunghi periodi di tempo fino a quando queste non si rivelano effettivamente necessarie, la rete neurale ricorrente viene modellata con celle LSTM e GRU.
Più nello specifico, si tratta di una rete neurale ricorrente molti a uno, con cinque strati, ovvero quattro hidden layers, di cui due composti da celle LSTM e due da celle GRU, e un output layer. L'input layer, come esplicitato nel paragrafo 3.3.2, non va considerato quando si definisce una rete neurale facendovi riferimento tramite il numero di strati da cui è composta. La numerosità degli hidden layers è stata scelta pari a quattro per modellare una rete neurale che fosse profonda, ma che non influisse in maniera eccessiva sulla complessità computazionale del modello.
Poiché, in generale, il numero di neuroni presenti negli hidden layers deve essere almeno pari alla dimensione dei dati di input, questo è stato posto pari a 400, in quanto ogni batch consiste in quattro caratteristiche di prezzo ripetute per cento giorni. Nell'output layer, invece, vi è un solo neurone, in quanto la previsione che la rete neurale deve effettuare riguarda il prezzo di chiusura di un singolo giorno.
Per ogni hidden layer viene aggiunto un tasso di dropout del 20%, come si può osservare dal codice in figura 52. Il dropout è una tecnica di regolarizzazione utilizzata per evitare il problema di overfitting, ovvero di sovra-adattamento della rete neurale che potrebbe semplicemente memorizzare i pattern contenuti nei dati di input, senza riuscire ad applicarvi la giusta

generalizzazione. Tramite il dropout, metodo economico dal punto di vista computazionale e notevolmente efficace, si migliora l'errore di generalizzazione commesso dalla rete neurale, escludendo in maniera casuale, durante l'addestramento della rete, alcuni output del layer a cui il dropout viene applicato, in base al tasso di dropout impostato. Questa tecnica ha l'effetto di rendere il processo di addestramento più rumoroso, inducendolo a ricercare una generalizzazione dei pattern presenti nei dati senza avere a disposizione delle informazioni completamente trasparenti.

```
n_neurons = x_train.shape[1] * x_train.shape[2]

model = Sequential()
model.add(LSTM(n_neurons, return_sequences=True,
               input_shape=(x_train.shape[1], x_train.shape[2])))
model.add(Dropout(0.2))
model.add(LSTM(n_neurons, return_sequences=True))
model.add(Dropout(0.2))
model.add(GRU(n_neurons, return_sequences=True))
model.add(Dropout(0.2))
model.add(GRU(n_neurons, return_sequences=False))
model.add(Dropout(0.2))
model.add(Dense(1))
```

Figura 52: architettura delle rete neurale

4.1.3 ADDESTRAMENTO

Dopo aver definito l'architettura delle rete neurale, si procede con l'addestramento della stessa tramite la definizione dell'algoritmo di apprendimento per la minimizzazione dell'errore. Nel paragrafo 3.3.3 è stato definito l'algoritmo gradient descent, ma, per ridurre i tempi di calcolo, spesso viene utilizzata una variante di tale algoritmo denominata stochastic gradient descent. Con lo stochastic gradient descent algorithm, il gradiente della funzione di costo non viene calcolato sull'intero set di dati di training, ma vengono presi in considerazione uno alla volta dei mini-batch, compresi all'interno dei batch descritti in precedenza e definiti sulla base del parametro batch size, e per ognuno di essi viene calcolato il gradiente della funzione di costo e vengono aggiornati i pesi sinaptici. Una volta che il procedimento di calcolo del gradiente e di aggiornamento dei pesi è stato eseguito per ogni mini-batch, allora si dice che un'epoca, ovvero un'iterazione, si è conclusa. Questo algoritmo è estremamente utile per evitare che il processo di apprendimento rimanga bloccato ad un minimo locale della curva della funzione di costo, non permettendo un apprendimento adeguato.

Con il metodo compile si specificano l'algoritmo di ottimizzazione, basato sullo stochastic gradient descent, e la funzione di costo. Per il modello in questione, come algoritmo di ottimizzazione è stato scelto Adam, un metodo di stochastic gradient descent che si basa sulla stima adattiva dei momenti di primo e secondo ordine, e che, nel libro del 2014 intitolato "Adam: A Method for Stochastic Optimization" di Diederik Kingma, Jimmy Ba, viene definito computazionalmente efficiente, con poca richiesta di memoria, adatto a problemi con quantità elevate di dati e parametri, e dunque uno tra i migliori ottimizzatori del processo di apprendimento delle reti neurali. Il tasso di

apprendimento predefinito per tale metodo di ottimizzazione è pari a 0,001. Invece, la funzione di costo scelta è quella che maggiormente viene utilizzata mei modelli di reti neurali, ovvero il Mean Squared Error (MSE), definito come la somma dei quadrati della differenza tra le variabili calcolate ed effettive divisa per il numero di elementi analizzati.
Con il metodo fit si definiscono i parametri su cui si sviluppa l'addestramento. Durante l'addestramento, la rete neurale elabora i dati contenuti in x_train e confronta le previsioni ottenute con i prezzi effettivi contenuti in y_train, andando poi a modificare i pesi associati ai neuroni della rete per minimizzare la funzione di costo. Il vettore y_train comprende i dati target che il modello cerca di prevedere tramite l'elaborazione dei prezzi contenuti in x_train e tramite la definizione della funzione di mappatura. Con il parametro batch size, come specificato in precedenza, si definisce il numero di osservazioni presenti in ogni mini-batch, ovvero il numero di campioni necessari per l'aggiornamento iterativo del gradiente, ed è stato impostato a 32, come da livello predefinito del metodo fit per la libreria Keras. Il numero di epoche, invece, è stato scelto pari a 15, ovvero il modello elaborerà per 15 volte tutti i dati compresi nel set di training, per completare l'addestramento.
Inoltre, viene utilizzato l'Early Stopping, una procedura che permette di interrompere l'addestramento prematuramente, ovvero senza completare le iterazioni definite dalle epoche, nel caso in cui il parametro loss, che definisce la minimizzazione della funzione di costo monitorando la qualità dell'addestramento, smetta di migliorare. Poiché l'obiettivo dell'addestramento risulta quello di minimizzare la funzione di costo, ad ogni ciclo di addestramento si controlla se alla fine dell'epoca in questione la loss sia diminuita o meno rispetto alla loss dell'epoca precedente, e, se tale requisito non risulta rispettato, allora, dopo aver atteso per un ulteriore numero di epoche definite dal parametro patience per assicurarsi che la loss non torni a diminuire, l'addestramento viene interrotto.
L'addestramento della rete neurale appena descritto è osservabile dal codice in figura 53.

```
model.compile(optimizer='adam', loss='mean_squared_error')
early_stop = EarlyStopping(monitor='loss', patience=5, verbose=1)
history = model.fit(x_train, y_train, batch_size=32,
                    epochs=15, callbacks=[early_stop])
```

Figura 53: addestramento della rete neurale

La minimizzazione della funzione di costo durante l'addestramento della rete neurale è rappresentato in figura 54 per il bitcoin, in figura 55 per l'EuroStoxx 50 e in figura 56 per l'S&P 500. Per quanto riguarda le reti neurali relative a bitcoin e ad EuroStoxx 50 l'addestramento ha avuto luogo fino all'ultima epoca definita in fase di costituzione del modello, mentre per quanto riguarda la rete neurale relativa a S&P 500 l'addestramento si è interrotto all'undicesima epoca, in quanto il parametro loss della settima all'undicesima epoca è stato costantemente superiore rispetto a quello della sesta epoca e, dunque, dopo avere aspettato per cinque epoche, come determinato tramite il parametro patience, l'addestramento è stato arrestato anticipatamente.

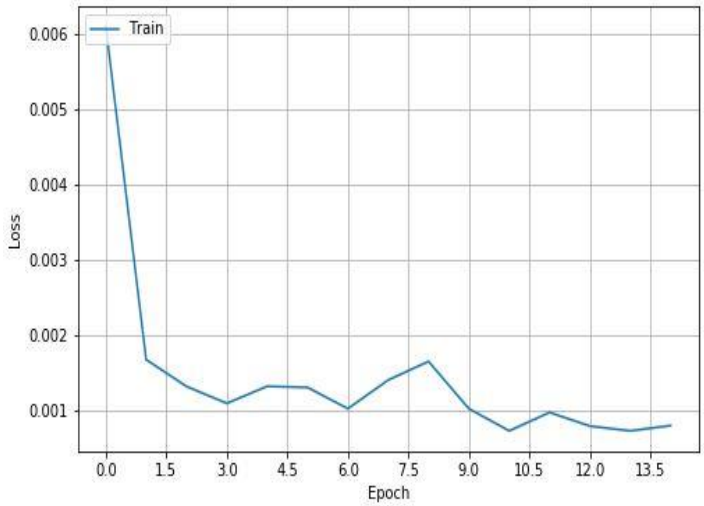

Figura 54: loss per la rete neurale di BTC/USD

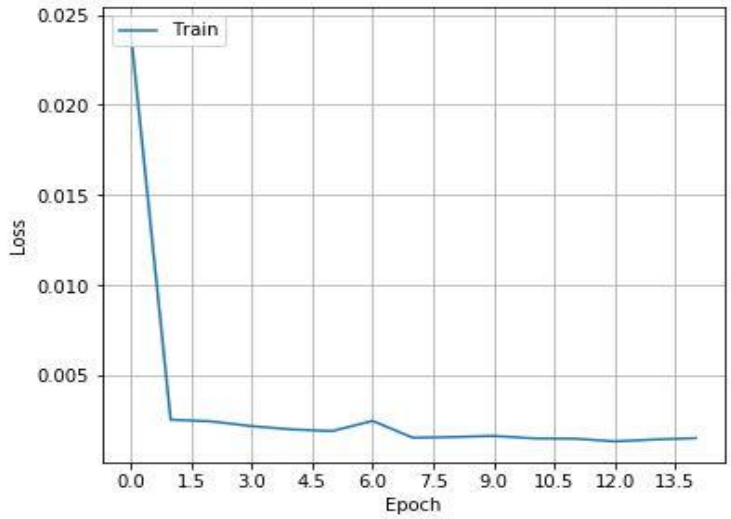

Figura 55: loss per la rete neurale di EuroStoxx 50

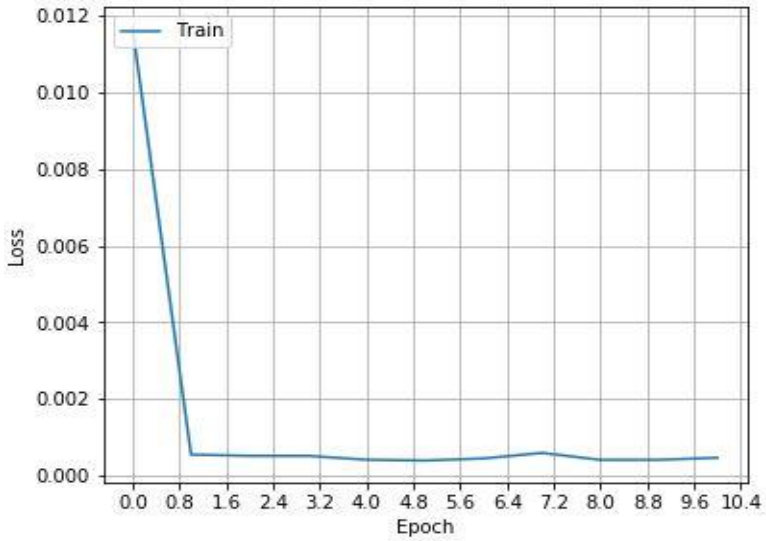

Figura 56: loss per la rete neurale di S&P 500

4.1.4 RISULTATI

Dopo aver effettuato l'addestramento, la rete neurale viene testata sui dati contenuti nell'array x_test del testing set per stabilire la capacità di generalizzazione della funzione di mappatura ricavata durante l'apprendimento supervisionato. Grazie al codice in figura 57, il modello elabora delle previsioni di prezzo ad un giorno per l'intervallo di tempo definito da y_test, il quale funge da metro di paragone con i risultati ottenuti dal modello per definire quanto le previsioni siano coincidenti o meno con i prezzi effettivi. Inoltre, poiché la rete neurale è stata addestrata con dati scalati attraverso la normalizzazione, allora anche le previsioni ottenute risultano comprese nell'intervallo di valori tra 0 e 1. Dunque, è necessario invertire l'operazione di normalizzazione effettuata in precedenza.

```
predictions = model.predict(x_test)
pred_unscaled = scaler_pred.inverse_transform(predictions)
```

Figura 57: previsioni del modello

Il confronto tra i prezzi previsti dal modello e i prezzi effettivi si può osservare nelle seguenti figure, dove in figura 58 vengono rappresentati i risultati del modello relativo al bitcoin, in figura 59 quelli del modello relativo all'EuroStoxx 50 ed infine in figura 60 quelli relativi all'S&P 500. Il grafico a linee mostra l'andamento dei prezzi previsti, indicati tramite la linea rossa, in relazione all'andamento dei prezzi effettivi, indicati tramite la linea nera. Invece l'istogramma, sottostante al grafico precedentemente descritto, rappresenta le differenze giornaliere tra prezzo previsto ed effettivo: il colore verde delle colonne riflette una differenza positiva e dunque una previsione ottimistica, mentre il colore rosso denota una differenza negativa e dunque una previsione pessimistica.
Una caratteristica comune tra i tre modelli risulta essere la risposta poco reattiva a variazioni significative di prezzo, sebbene il modello relativo all'EuroStoxx 50 sia quello che meglio si adatta a tali movimenti di prezzo. Generalmente, quando il prezzo assume un trend di breve periodo rialzista, i modelli tendono a generare previsioni pessimistiche, mentre avviene il contrario quando il prezzo assume un trend di breve periodo ribassista. Invece, quando il prezzo si assesta attorno ad una situazione di equilibrio, seppur temporanea, allora i modelli riescono ad ottenere delle previsioni abbastanza accurate.
Sebbene, a primo impatto, osservando le figure appena menzionate, il modello relativo al bitcoin possa sembrare quello maggiormente accurato per una previsione corretta dei prezzi, è necessario verificare le prestazioni dei modelli tramite delle metriche di errore che non siano affette dalla larghezza dell'intervallo di valori entro cui variano i prezzi. Infatti, il prezzo di bitcoin, nel periodo di validazione del modello, ha un minimo al valore di $4857,10 e un massimo di $19172,84, rivelando così una variazione di prezzo considerevolmente più ampia rispetto a quella che si verifica nei due indici azionari in analisi, EuroStoxx 50 e S&P 500, i quali, rispettivamente, variano tra $2375,70 e $3856,20 e tra $2237,40 e $3635,41. L'ampia variazione a cui è soggetto il prezzo di bitcoin non permette di percepire attraverso la sola osservazione del grafico in figura 57 la significatività delle differenze tra prezzi previsti ed effettivi.

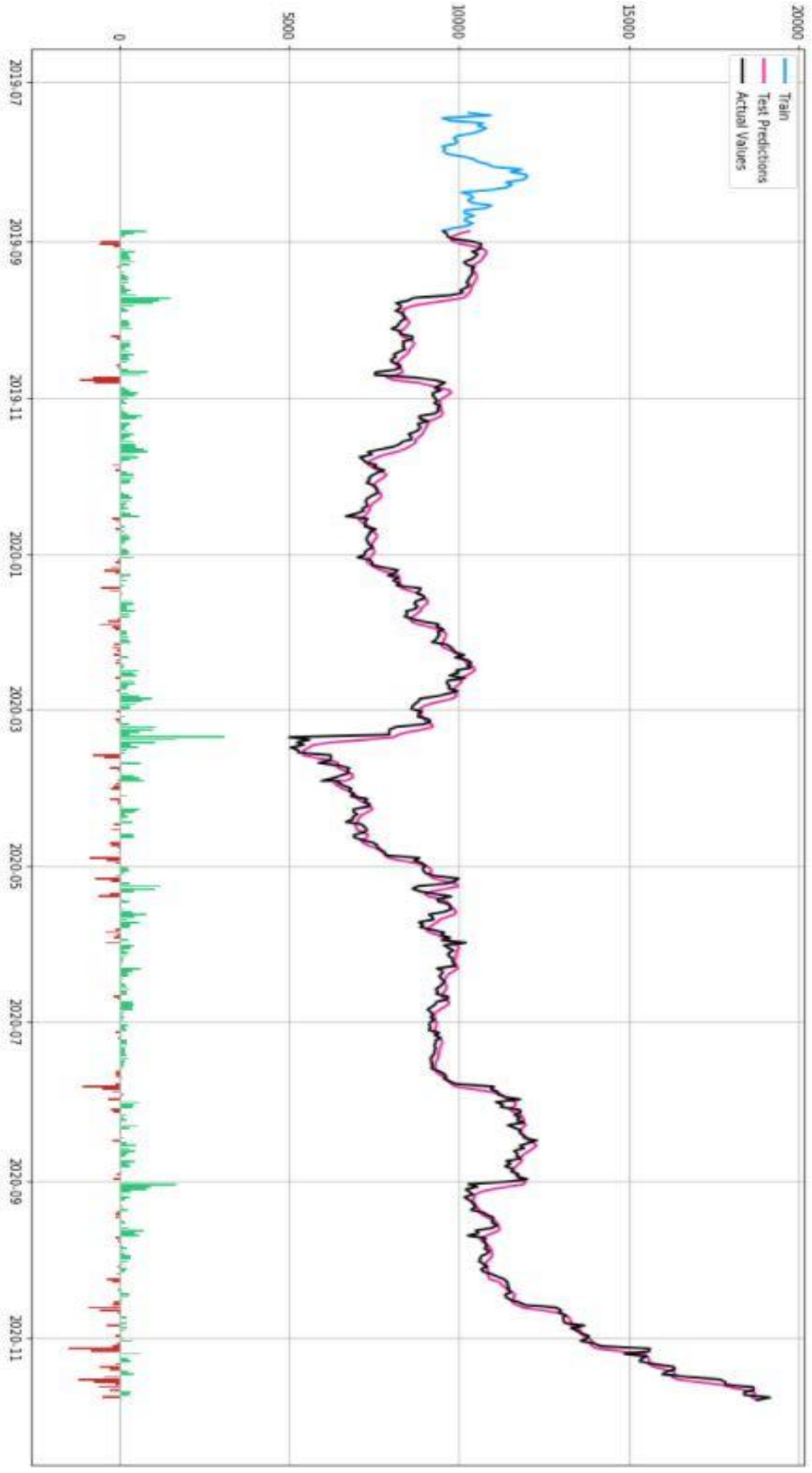

Figura 58: risultati modello BTC/USD

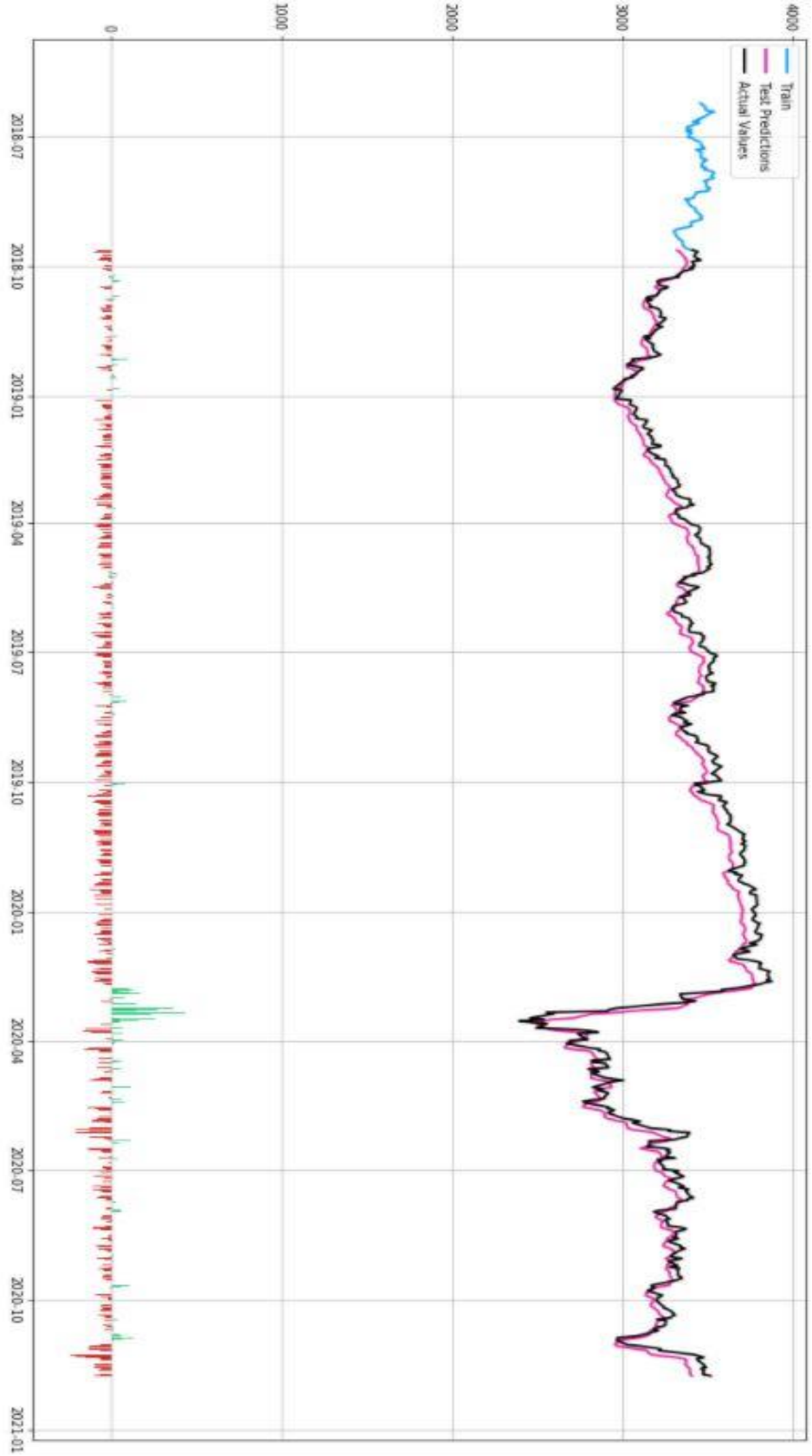

Figura 59: risultati modello EuroStoxx 50

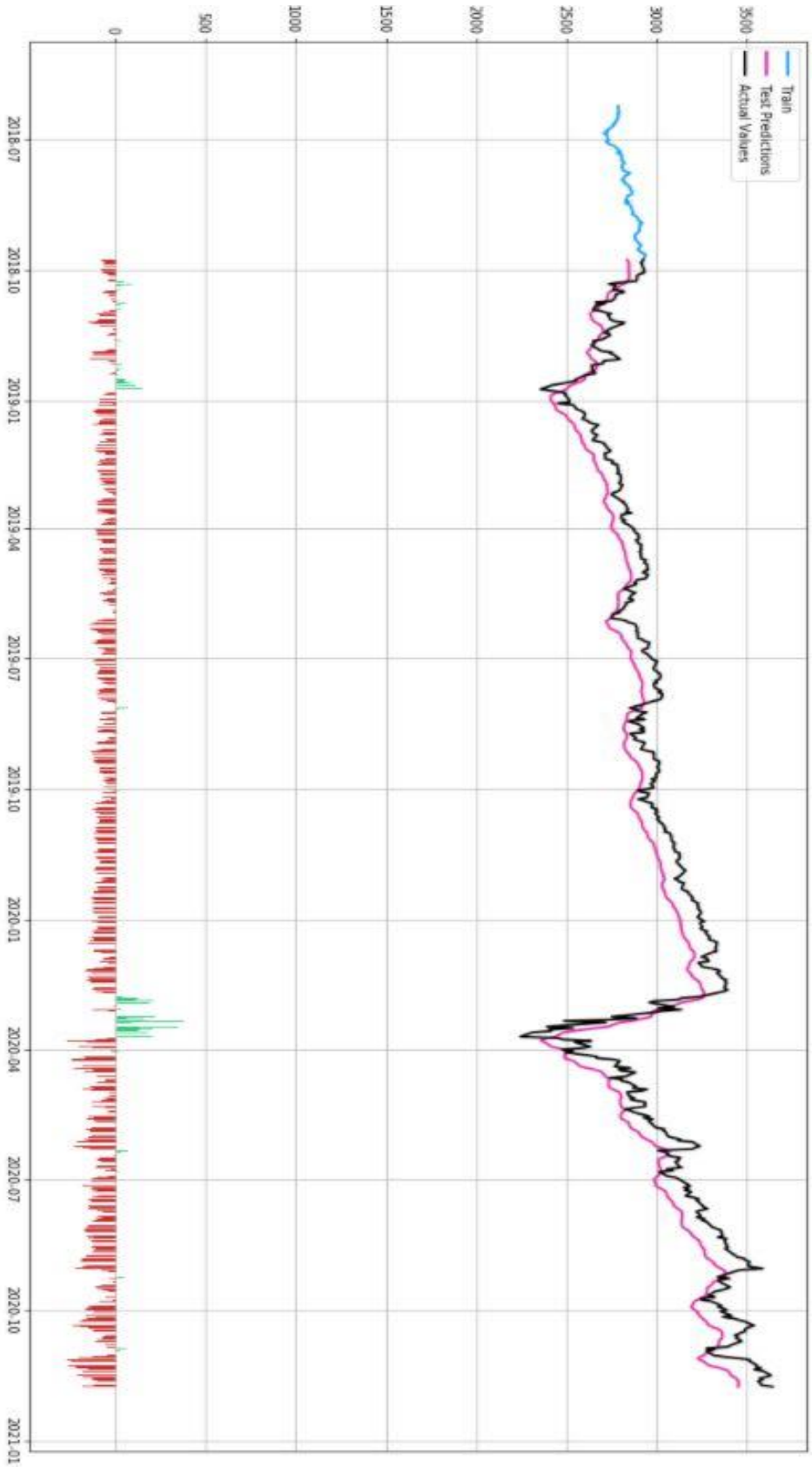

Figura 60: risultati modello S&P 500

Nella seguente tabella vengono riportati gli errori commessi dalla rete neurale sui modelli relativi a bitcoin, EuroStoxx 50 e S&P 500, misurati secondo le metriche di mean absolute error (MAE), median absolut error (MDAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) e median absolute percentage error (MDAPE).

	MAE	MDAE	RMSE	MAPE	MDAPE
BTC/USD	281,39	214,5	118,87	3,17%	2,16%
EUROSTOXX 50	65,99	61,32	47,85	2,0%	1,82%
S&P500	103,53	103,93	89,91	3,43%	3,5%

Tabella 1: prestazioni dei modelli BTC/USD, EuroStoxx 50 e S&P 500

Si può osservare come il MAE, il MDAE e l'RMSE del modello relativo a bitcoin siano di gran lunga maggiori rispetto a quelli ottenuti con i modelli relativi ai due indici azionari. Dalla relativa somiglianza tra gli errori misurati tramite il MAE e il MDAE, per tutti e tre gli asset considerati, si può evincere che non vi siano particolari valori estremi, definiti outliers, che influenzano negativamente le prestazioni dei modelli, e tale risultato viene confermato dall'RMSE.
Per la comparazione tra i modelli risulta più conveniente soffermarsi sulle ultime due metriche. Il MAPE rivela che le previsioni effettuate sul prezzo del bitcoin sono realizzate con un errore medio del 3,17%, mentre quelle sull'EuroStoxx 50 ammettono un errore medio del 2,0% e quelle sull'S&P 500 sono associate ad un errore medio del 3,43%. I risultati dedotti dal MAPE vengono avvalorati dalla metrica successiva, in quanto dall'osservazione dei valori riportati nell'ultima colonna della tabella, per quanto riguarda il bitcoin, l'errore mediano commesso risulta pari al 2,16%, mentre, per quanto riguarda l'EuroStoxx 50 e l'S&P 500, gli errori mediani sono del 1,82% e del 3,5% rispettivamente.
Al fine di riuscire a comprendere il significato di questi dati bisogna fare una riflessione relativa alle variazioni percentuali giornaliere dei prezzi degli asset considerati. Nel periodo analizzato, la variazione percentuale giornaliera media, calcolata in valore assoluto, per i prezzi del bitcoin è pari a 3,52%, mentre per l'EuroStoxx 50 e l'S&P 500 essa è dello 0,89% e dello 0,71% rispettivamente. Tale misura rivela l'elevata volatilità che si riscontra nel mercato delle criptovalute, il quale registra variazioni percentuali giornaliere che in alcuni casi possono giungere anche a valori superiori del 20%. Alla luce di tale osservazione, si può capire come un errore medio di previsione del prezzo futuro pari a 3,17%, se considerato in riferimento ad una variazione giornaliera media del 3,52%, sia nettamente più tollerabile e ragionevolmente accettabile rispetto ad un errore di previsione del 2% o del 3,43%, i quali sono rispettivamente da considerare in riferimento ad una variazione giornaliera media dello 0,89% e dello 0,71%.

Dunque, a fronte di una variazione giornaliera maggiore, l'errore commesso dal modello relativo al bitcoin rivela una previsione più accurata della quotazione futura del prezzo, rispetto a quella ottenuta per i due indici azionari. Poiché, come esplicitato in precedenza, in un mercato efficiente dovrebbe essere impossibile stimare in modo sistematico l'andamento dei prezzi e la conoscenza dei prezzi passati non dovrebbe essere utile a predire i prezzi futuri, allora, per le considerazioni appena espresse, il mercato delle criptovalute risulta meno efficiente del mercato azionario.
Per un ulteriore analisi, al fine di comprendere l'evoluzione avvenuta nel mercato delle criptovalute, è stato diviso il set di dati disponibili sul bitcoin in due parti, una che comprende gli anni dal 2014 al 2017 e una dal 2017 al 2020, e questi due dataset sono stati immessi come input nella rete neurale, mantenendo la stessa architettura descritta in precedenza.
Come si può notare sia dalle figure 61 e 62, che dalla tabella 2, nel primo periodo considerato, valutando gli errori in riferimento alla variazione percentuale giornaliera media degli asset in analisi, l'errore di previsione commesso, in termini percentuali, risulta proporzionale a quello osservato nei modelli degli indici azionari, mentre nel secondo periodo i risultati ottenuti sono comparabili a quelli conseguiti nel modello che prende in considerazione il periodo che va dal 2014 al 2020. Dunque, si può concludere affermando che l'efficienza del mercato delle criptovalute si è ridotta a partire dalla fine del 2017, periodo corrispondente allo scoppio della bolla speculativa avvenuta su tale mercato.

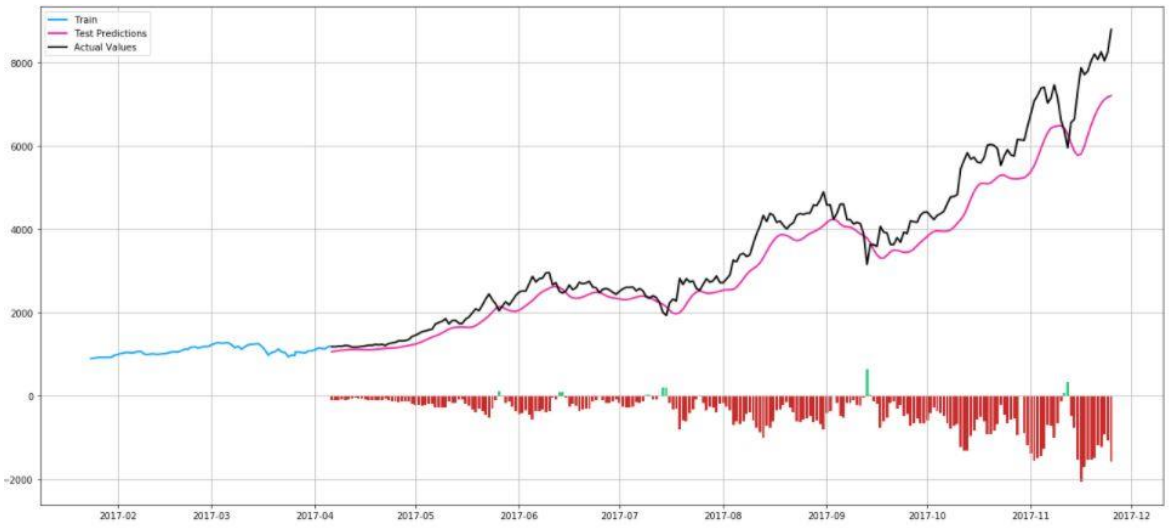

Figura 61: risultati modello BTC/USD nel periodo 2014-2017

Figura 62: risultati modello BTC/USD nel periodo 2017-2020

	MAE	MDAE	RMSE	MAPE	MDAPE
BTC/USD 2014-2017	440,65	326,1	425,2	11,32%	11,11%
BTC/USD 2017-2020	372,55	265,21	310,51	3,25%	2,63%

Tabella 2: prestazione dei modelli BTC/USD 2014-2017 e BTC/USD 2017-2020

5. CONCLUSIONE

Il mercato delle criptovalute è sicuramente meno maturo rispetto al mercato azionario e dalla comparazione, tramite la rete neurale, del grado di efficienza tra di essi è stato dimostrato come il primo sia meno efficiente del secondo, in quanto è possibile prevedere con maggiore accuratezza quale sarà la quotazione futura, cosa che secondo l'EMH non dovrebbe avvenire. Inoltre, dalla successiva analisi eseguita solamente sui prezzi del bitcoin, si è potuto osservare il fatto che l'efficienza sia diminuita con il tempo e che dunque essa non dipenda dalla maturità del mercato, almeno per quanto riguarda quello delle criptovalute.

Per un ulteriore verifica di quanto affermato, viene di seguito illustrato brevemente il modello Stock-to-Flow (S2F) ideato da PlanB, pseudonimo dietro al quale si cela un investitore istituzionale olandese con un background legale e di finanza quantitativa e che attualmente lavora come responsabile degli investimenti in un team che gestisce un bilancio multimiliardario.

Il modello S2F è stato pubblicato a marzo 2019 sulla piattaforma Medium, come un modello di valutazione di bitcoin, ispirato all'analisi compiuta da Nick Szabo sul concetto di onerosità non falsificabile e sugli studi compiuti sul rapporto stock-to-flow (SF) da Saifedean Ammous, autore del libro "The Bitcoin Standard: The Decentralized Alternative to Central Banking", pubblicato nel 2018.

Il rapporto SF, che può essere calcolato anche per altri beni come ad esempio l'oro, è una misura della scarsità del bene in questione, in quanto lo stock misura la dimensione delle scorte o delle riserve esistenti, mentre il flow rappresenta la produzione annuale. Dal grafico in figura 63 si può notare come un dimezzamento del flow, definito halving, ogni 210.000 blocchi processati, ovvero circa ogni quattro anni, porti ad un inflazione monetaria del bitcoin decrescente e al raggiungimento del massimo stock possibile, definito in 21 milioni di bitcoin, con un movimento parabolico asintotico.

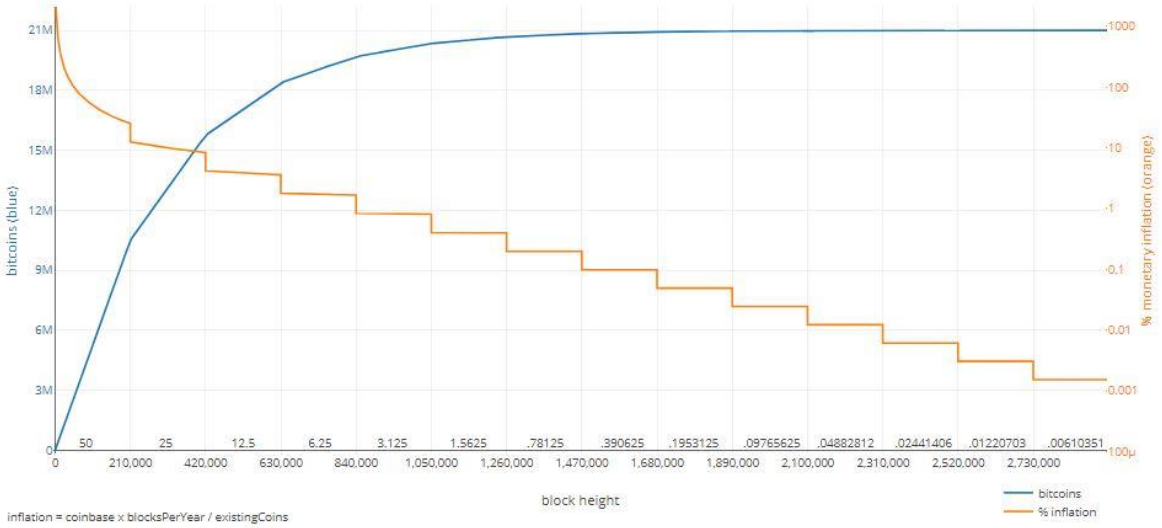

Figura 63: inflazione monetaria del bitcoin

Tale struttura della produzione e dell'offerta del bitcoin porta ad un SF che raddoppia ogni 210.000 blocchi. Il rapporto SF viene spesso calcolato secondo la seguente formula

$$SF = 1/tasso\ di\ inflazione\ monetaria\ di\ bitcoin$$

Il modello S2F è una legge di potenza che è stata calcolata sui dati mensili rilevati nell'intervallo di tempo tra ottobre 2009 e febbraio 2019, che mette in relazione il rapporto SF con la capitalizzazione di mercato del bitcoin, tramite la funzione di regressione lineare

$$\ln(capitalizzazione\ di\ mercato) = 3.3 * \ln(SF) + 14.6$$

la quale può essere espressa come una funzione di legge di potenza, ovvero

$$capitalizzazione\ di\ mercato = e^{[(3.3*\ln(SF))+14.6]}$$

Una legge di potenza è una relazione in cui una variazione relativa di una quantità provoca una variazione relativa proporzionale nell'altra quantità, indipendentemente dalla dimensione iniziale di tali quantità. In occasione di ogni halving, il rapporto SF di bitcoin raddoppia e il valore di mercato aumenta di dieci volte, e questo è un fattore costante. Inoltre, le leggi di potenza sono interessanti perché rivelano una regolarità di fondo nelle proprietà di sistemi complessi apparentemente casuali.
Tale relazione dà origine ad una previsione del prezzo futuro del bitcoin che può essere sintetizzata con la seguente formula

$$prezzo\ BTC = 0.4 * SF^3$$

Nel grafico in figura 64 viene presentato il modello di previsione, dove tramite la linea blu si delineano le quotazioni del bitcoin previste dal modello, aggiungendo anche un'area di incertezza in azzurro per la volatilità di breve periodo che caratterizza il bitcoin, e dove tramite i pallini arancioni vengono rappresentati i prezzi di chiusura mensili del bitcoin su scala logaritmica.

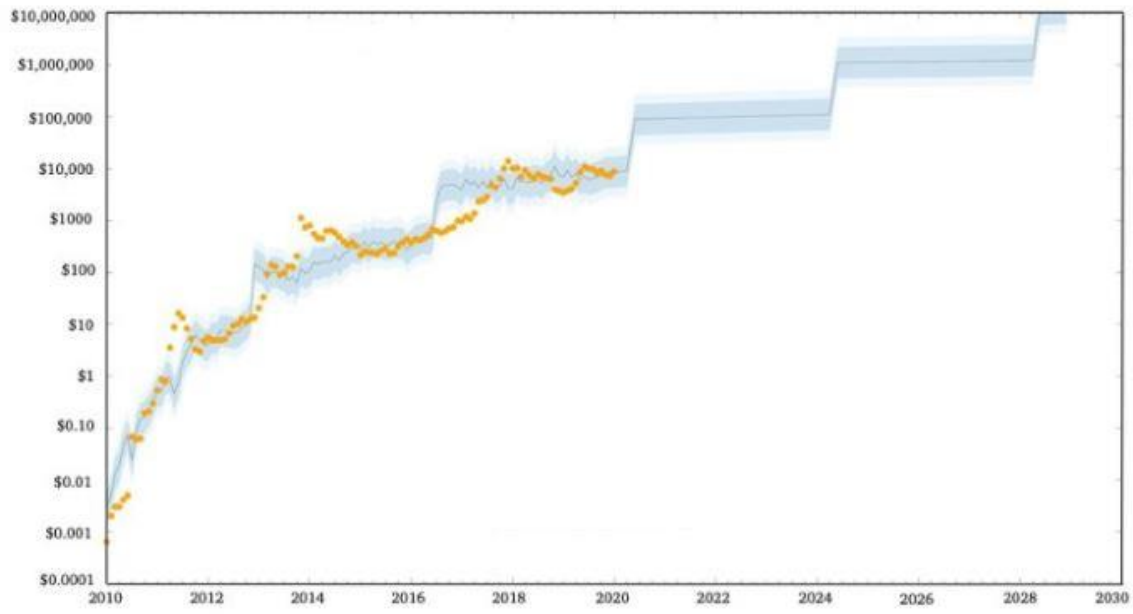

Figura 64: modello S2F

La relazione, regolata dalla legge di potenza, tra il modello S2F e il prezzo del bitcoin mostra l'esistenza di una regolarità sottostante il complesso sistema dinamico del bitcoin, la cui quotazione tende a stringersi sempre di più tra le aree di scostamento teorizzate. Grazie a questo modello,

ideato a marzo 2019, è stata formulata una previsione per cui il prezzo di bitcoin sarebbe diventato uno o due anni dopo l'halving del 2020 pari a $ 55.000.
Tale modello contrasta con quanto affermato dall'ipotesi di efficienza dei mercati, secondo la quale le informazioni su cui si basa il modello, poiché sono pubblicamente disponibili, dovrebbero già essere incluse nei prezzi, così come i risultati del modello. Però, questo non avviene, come si può notare dall'avverarsi della previsione ipotizzata ad inizio 2019, realizzatasi a fine febbraio 2021. Inoltre, la regolarità con cui avvengono le variazioni successive dei prezzi, come osservato in figura 63, rappresenta esattamente l'opposto della casualità con cui, in un mercato che elabora in modo efficiente le informazioni disponibili agli operatori, dovrebbero verificarsi gli scostamenti dai prezzi precedenti.
Dunque, tale modello, che si basa su dati di natura fondamentale piuttosto che sui prezzi passati, avvalora i risultati ottenuti nel capitolo precedente con la rete neurale.
Una possibile spiegazione alla riscontrabile inefficienza informativa del bitcoin e quindi del mercato delle criptovalute può essere l'eccessiva considerazione che viene attribuita ad alcuni rischi che incombono su tale mercato, come ad esempio una regolamentazione severa da parte dei governi, che potrebbero bandire lo scambio di alcune criptovalute rendendole illegali, un 51% attack al network di una determinata criptovaluta, un errore di programmazione del software, un attacco hacker agli exchange su cui vengono custodite le criptovalute, un hardfork, o infine una rivolta dei miners che dopo gli halving vedono la loro ricompensa per la validazione dei blocchi dimezzata.
Tutti questi rischi concorrono alla non incorporazione delle informazioni pubblicamente disponibili nei prezzi.

BIBLIOGRAFIA E SITOGRAFIA

Slides, Analisi Finanziaria e Creditizia per l'Impresa, Politecnico di Torino, a.a. 2019/2020, Prof. Franco Varetto

AI Sciences, "Deep Learning: Recurrent Neural Network with Python", Udemy, 2020

Bachelier L., "Théorie de la spéculation", Annales scientifiques de l'École Normale Supérieure, 1900

Bank for Internation Settlements, "Impending arrival – a sequel to the survey on central bank digital currency", 01/2020

Bernstein P., "Against the Gods: The Remarkable Story of Risk", 1998

Black F., "Noise", Journal of Finance, 1986

Bryson A., Ho Y., "Applied optimal control: optimization, estimation, and control", 1969

Campbell M., Hoane J., Hsu F., "Deep Blue", Elsevier, 2002

Carhart M., "On Persistence in Mutual Fund Performance", Journal of Finance 52, 1997

Chaum D., "Advances in Cryptology: Proceedings of Crypto 82", 1983

Cho K., Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", 2014

Committee on Payments and Market Infrastructures, "Digital Currencies", 2015

Cybenko G., "Approximation by superpositions of a sigmoidal function", Mathematics of Control, Signals, and Systems, 1989

DeBondt W., Thaler R., "Does the Stock Market Overreact?", Journal of Finance 40, 1985

European Banking Authority, "EBA Opinion on 'virtual currencies'", 04/07/2014

European Central Bank, "Crypto-Assets: Implications for financial stability, monetary policy, and payments and market infrastructures", 2019

European Central Bank, "Report on a digital euro", 10/2020

European Central Bank, "Virtual Currency Schemes", 2012

European Central Bank, "Virtual Currency Schemes – a further analysis", 2015

European Parliament, “Cryptocurrencies and blockchain - Legal context and implications for financial crime, money laundering and tax evasion”, 07/2018

European Securities and Markets Authority, “WARNING - ESMA, EBA and EIOPA warn consumers on the risks of Virtual Currencies”, 12/02/2018

Fama E., “The Behavior of Stock Market Prices”, Journal of Business 38, 1965

Fama E., “Random Walks in Stock Market Prices”, Financial Analysts Journal 21, 1965

Fama E., “Efficient Capital Markets: A Review of Theory and Empirical Work”, Journal of Finance 25, 1970

Fama E., Fisher L., Jensen M., Roll R., “The Adjustment of Stock Prices to New Information”, International Economic Review, 1969

Fama E., French K., “Permanent and Temporary Components of Stock Prices”, Journal of Political Economy 96, 1988

Fama E., French K., “The Cross-Section of Expected Stock Returns", Journal of Finance 47, 1992

Fama E., French K., “Common Risk Factors in the Returns on Stock and Bonds", Journal of Finance 33, 1993

Federal Reserve Bank of St. Louis, “The Case for Central Bank Electronic Money and the Non-case for Central Bank Cryptocurrencies”, 2018

Financial Action Task Force, “Virtual Currencies - Key Definitions and Potential AML/CFT Risks”, 06/2014

Fluck Z., Malkiel B., Quandt, R., “The Predictability of Stock Returns: A Cross-Sectional Simulation” Review of Economics and Statistics 79, 1997

Graham B., Dodd D., “The Intelligent Investor”, 1965

Goodfellow I., Bengio Y., Courville A., “Deep Learning”, 2016

Goodfellow I., Bulatov Y., Ibarz J., Arnoud S., Shet, V., “Multi-digit number recognition from Street View imagery using deep convolutional neural networks”, International Conference on Learning Representations, 2014

Grossman S., Stiglitz J., “On the Impossibility of Informationally Efficient Markets”, The American Economic Review 70, 1980

Guerini M., “Intelligenza Artificiale, forte e debole, tra mito e realtà”, Corriere della sera Blog, 2017

Haugen R., Lakonishok J., “The Incredible January Effect: The Stock Market’s Unsolved Mistery”, 1987

Hinton G., Osindero S., Teh Y., “A fast learning algorithm for deep belief nets”, Neural Computation, 2006

Hinton G., Salakhutdinov R., “Reducing the dimensionality of data with neural networks”, Science, 2006

Hochreiter S., Schmidhuber J., “Long Short-Term Memory”, Neural Computation, 1997

International Monetary Fund “Virtual Currencies and Beyond: Initial Considerations”, 2016

Jensen M., “The Performance of Mutual Funds in the Period 1945–1964”, Journal of Finance, 1968

Keynes J., “The General Theory of Employment, Interest and Money”, 1936

Kingma D., Ba J., "Adam: A Method for Stochastic Optimization" 2014

Lo A., MacKinlay C., “A Non-Random Walk Down Wall Street”, 1999

Malkiel B., “The Efficient-Market Hypothesis and the Financial Crisis”, 2011

Malkiel B., “A Random Walk Down Wall Street”, 1973

McClelland J., Rumelhart D., and Hinton G., “The appeal of parallel distributed processing”, Computation & intelligence, 1995

McCulloch W., Pitts W., “A logical calculus of the ideas immanent in nervous activity”, Bulletin of Mathematical Biophysics, 1943

Miller M., “Financial Innovations and Market Volatility”, 1991

Minsky L., Papert A., “Perceptrons”, 1969

Mitchell T., “Machine Learning” 1997

Mitchell T., “The Discipline of Machine Learning”, 2006

O’Higgins M., “Beating the Dow”, 1991

Roberts H., “Stock Market "Patterns" and Financial Analysis: Methodological Suggestions.”, Journal of Finance 14, 1959

Roll R., “Volatility of U.S. and Japanese Stock Markets: A Symposium” Journal of Applied Corporate Finance, 1992

Rosenblatt F., “The perceptron: A probabilistic model for information storage andorganization in the brain”, Psychological Review, 1958

Rumelhart D., Hinton G., Williams R., “Learning representations by back-propagating errors”, Nature, 1986

Russell S., Norving P., “Artificial Intelligence: a modern approach” 2005

Searle J., “Minds, brains, and programs”, Behavioral and Brain Sciences, 1980

Shiller R., “Irrational Exuberance”, 2000

Solnik B., “A Note on the Validity of the Random Walk for European Stock Prices.” Journal of Finance 28, 1973

Van Veen F., Leijnen S., “The Neural Network Zoo”, Asimov Institute, 2019

Widrow B., Hoff E., “Adaptive switching circuits”, IRE WESCON Convention Record, 1960

World Bank, “Distributed Ledger Technology (DLT) and blockchain”, 2017

World Economic Forum, “Central Banks and Distributed Ledger Technology: How are Central Banks Exploring Blockchain Today?”, 03/2019

Wright R., “Can machines think?”, Time Magazine, 1996

https://www.bancaditalia.it/

https://www.bitcoin.org/

https://www.blockchain4innovation.it/

https://www.bloomberg.com/

https://www.borsaitaliana.it/

https://www.coindesk.com/

https://www.coinmarketcap.com/

https://www.cointelegraph.com/

http://www.consob.it/

https://www.cryptonomist.ch/

https://www.digitaldollarproject.org/

https://www.ecb.europa.eu/

https://www.geeksforgeeks.org/

https://www.keras.io/

https://www.imf.org/

https://www.investopedia.com/

https://www.medium.com/

https://www.startingfinance.com/

https://towardsdatascience.com/

https://www.weidai.com/bmoney.txt