

Covid e FCCR: classificazione con CNN

Gennaro Capaldo, Carmela Pia Senatore





Di cosa tratteremo?

Metodologia utilizzata

Descrizione dello stato dell'arte

1

Introduzione

Descrizione del
contesto biologico
e bioinformatico

2

Presentazione dei risultati

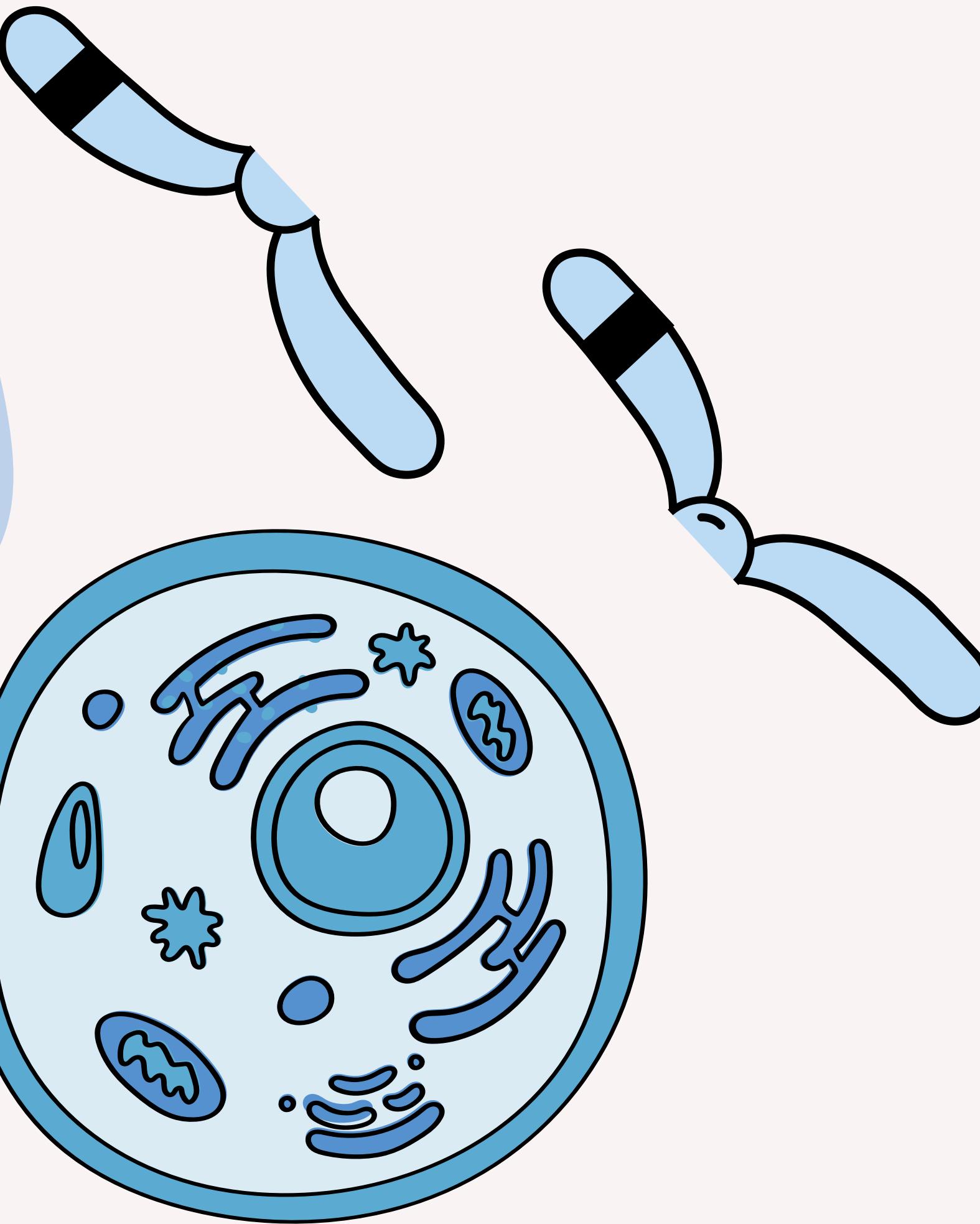
Analisi dei dati e
discussione dei risultati

3

Conclusioni

4

01. INTRODUZIONE



Covid-19: che cos'è?



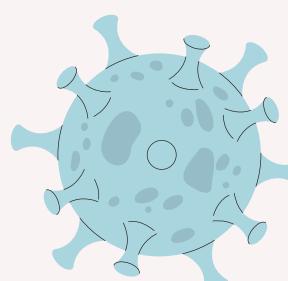
Fin dal suo primo rilevamento a Wuhan, in Cina, nel dicembre 2019, il virus si è diffuso rapidamente a livello globale, dando origine a una crisi sanitaria senza precedenti.



Il genoma del SARS-CoV-2 è costituito da circa 30.000 basi di RNA, che codificano per diverse proteine strutturali e non strutturali.



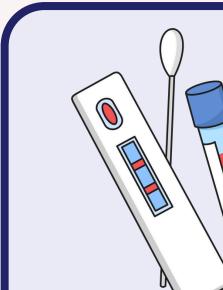
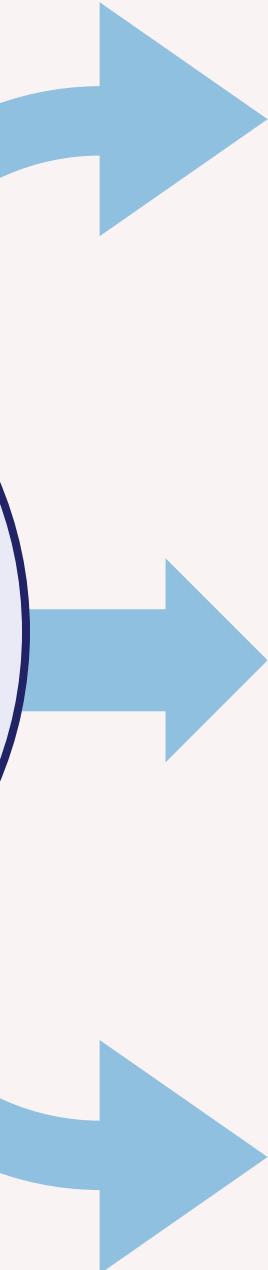
Il virus viene trasmesso principalmente tramite droplet e aerosol da una persona infetta quando starnutisce, tossisce, parla o respira e si trova in prossimità di altre persone.



Come ogni virus a RNA, il SARS-CoV-2 è soggetto a mutazioni che possono alterare la sua sequenza genomica.

Alcune di queste mutazioni possono conferire al virus vantaggi evolutivi, come una maggiore trasmissibilità o una parziale resistenza alla risposta immunitaria.

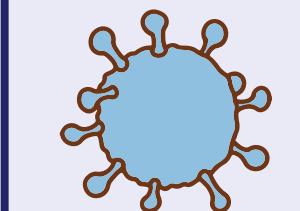
Classificazione delle varianti



VARIANTI DI INTERESSE: sono varianti che presentano mutazioni con caratteristiche genetiche che potrebbero influenzare la trasmissibilità, la gravità della malattia, l'efficacia dei vaccini o la risposta ai trattamenti.



VARIANTI DI PREOCCUPAZIONE: presentano evidenze di una maggiore trasmissibilità, una severità della malattia o una ridotta capacità di neutralizzazione da parte di anticorpi derivanti da infezioni precedenti o dalla vaccinazione.



VARIANTI SOTTO MONITORAGGIO: varianti con mutazioni che potrebbero rappresentare un rischio, ma per le quali l'evidenza scientifica non è ancora sufficiente per classificarle come VOI o VOC.

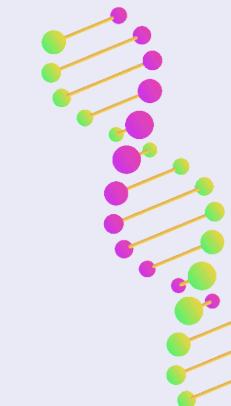


Identificazione del problema



CLASSE O

Raggruppa le varianti che non rientrano nelle categorie VOI o VOC identificate dalle autorità sanitarie. Questa classe include le varianti che sono rimaste relativamente stabili o che presentano mutazioni che non hanno avuto un impatto significativo sulla trasmissibilità.

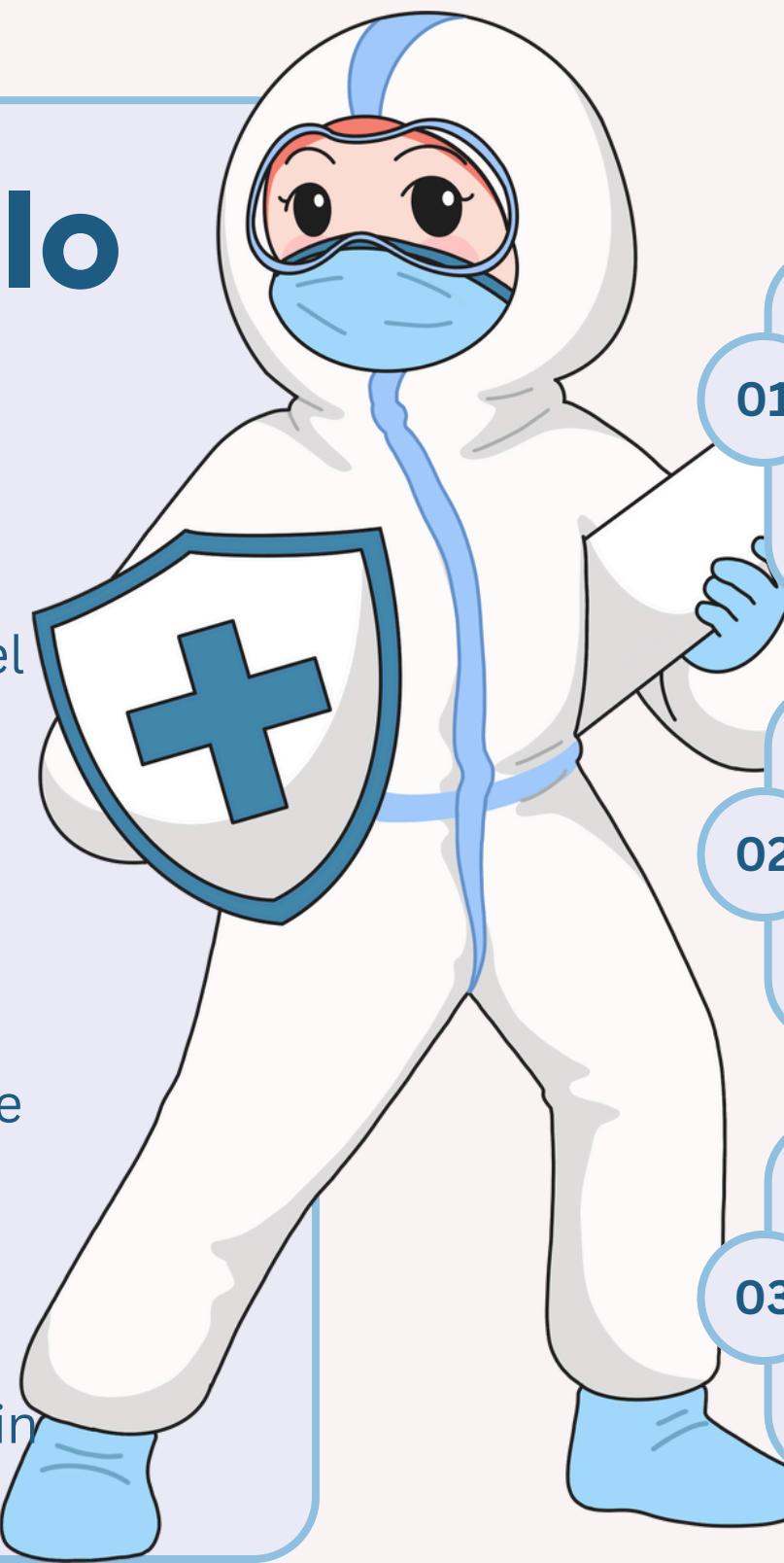


CLASSE S

Si concentra sulle varianti che presentano mutazioni significative nella proteina Spike (S), la proteina che media l'ingresso del virus nelle cellule ospiti attraverso il legame con il recettore ACE2. La proteina Spike è anche il principale bersaglio degli anticorpi neutralizzanti generati dall'infezione naturale e dalla vaccinazione, nonché il focus di molti vaccini contro il COVID-19.

Obiettivi dello studio

Tra le proteine, **Spike** (S) è di particolare interesse, in quanto è il principale mediatore dell'ingresso del virus nelle cellule umane. Il sequenziamento genomico è lo strumento principale utilizzato per identificare e classificare le varianti. Questo processo coinvolge la decodifica dell'intero genoma virale, permettendo agli scienziati di rilevare le mutazioni specifiche che caratterizzano ogni variante. Le sequenze genomiche ottenute vengono poi confrontate con sequenze di riferimento e archiviate in database globali come GISAID.



01

Analizzare le sequenze genomiche del SARS-CoV-2 appartenenti alle classi O e S per ottenere una comprensione più approfondita delle loro caratteristiche genetiche e funzionali

02

Per ogni paziente si otterrà una rappresentazione FCGR, con $k=7$.

03

Classificazione tramite rete neurale convoluzionale (CNN).

02. METODOLOGIA UTILIZZATA



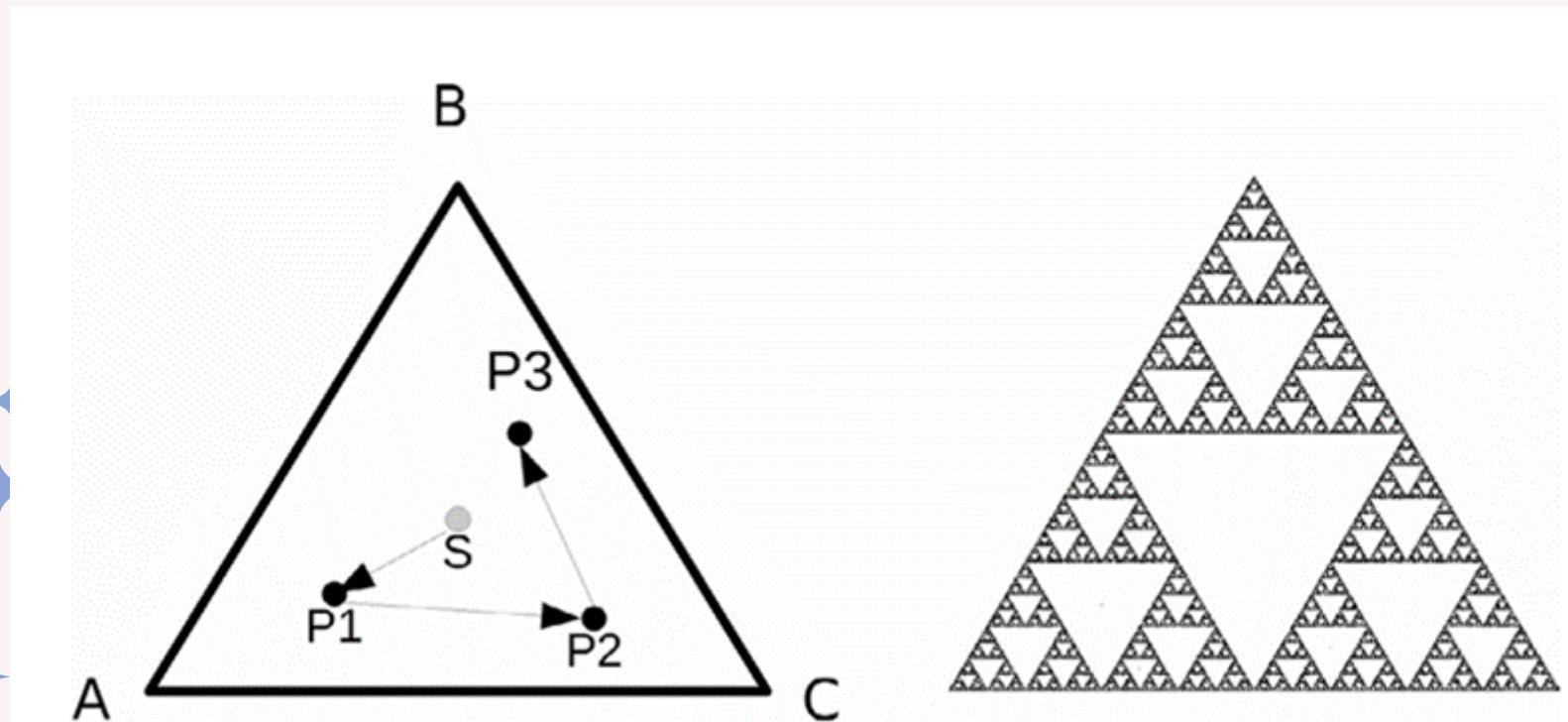


1. Chaos Game Representation

Barnsley e Jeffrey

La Chaos Game Representation (CGR) è una tecnica di mappatura iterativa che elabora sequenze di unità , come i nucleotidi di una sequenza di DNA o gli amminoacidi in una proteina, per trovarne le coordinate e la loro posizione in uno spazio continuo. Questa distribuzione di posizione è unica e la sequenza fonte può essere recuperata dalle coordinate in modo tale che la distanza tra le posizioni misura la somiglianza tra le sequenze corrispondenti.

Sistema di funzioni iterate



01

Definizione dell'algoritmo IFS: Un sistema di funzioni iterate (IFS) utilizza mappature contrattate in uno spazio metrico, ciascuna con una probabilità associata. Per il Triangolo di Sierpinski, si considerano tre mappe w_1, w_2, w_3 , ciascuna con probabilità $1\backslash 3$.

02

Generazione dell'immagine frattale: Basandosi su un punto iniziale selezionato casualmente (S), un vertice viene scelto casualmente (V_1) e viene disegnato un punto P_1 a metà della distanza dal vertice V_1 . Questo processo viene ripetuto, con P_1 come nuovo punto di partenza. Questo processo viene ripetuto molte volte.

03

Visualizzazione del frattale: Dopo numerose iterazioni, i punti generati mostrano il pattern del frattale. Nel caso del Triangolo di Sierpinski, i punti convergono verso una figura con i vertici nei punti $(0,0)$, $(0,1)$, e $(1,1)$.

CGR AL DNA

Principali caratteristiche

Unicità

- una sequenza è rappresentata come un modello unico
- una sequenza viene mappata su coordinate univoche
- una singola coordinata codifica l'input della sequenza completa

Passaggio

- un CGR mappa tutte le possibili sequenze in tutte le possibili lunghezze in 2D o in uno spazio 3D

Randomicità

- il punto di partenza influenza di poco il risultato

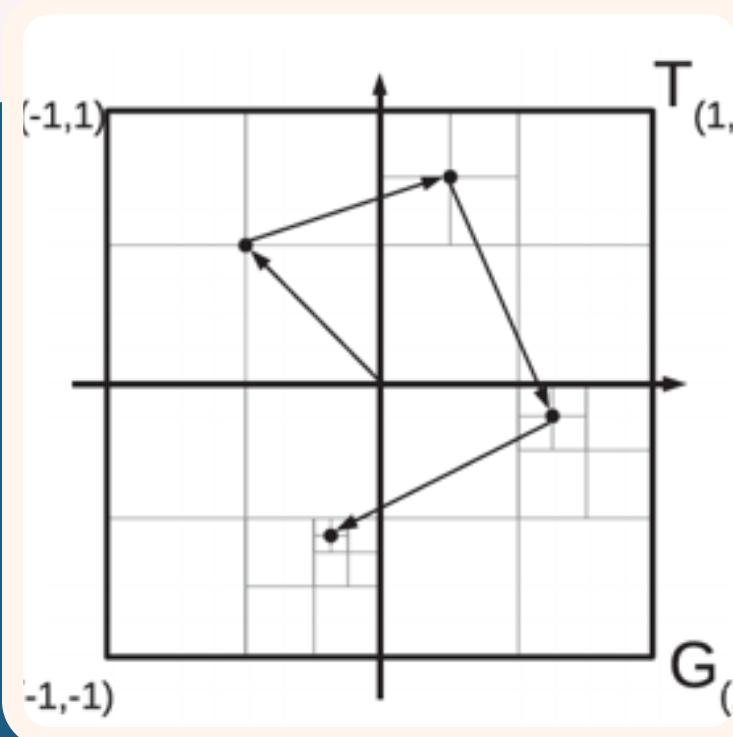
CGR AL DNA(2)

Invece di usare la rappresentazione a triangolo, il CGR era basato su un quadrato, con i quattro vertici che rappresentavano i quattro nucleotidi: adenina (A), citosina (C), guanina (G) e timina (T) o uracile (U) per DNA e RNA rispettivamente.

Jeffrey ha osservato che per sequenze casuali non emergono modelli visibili.

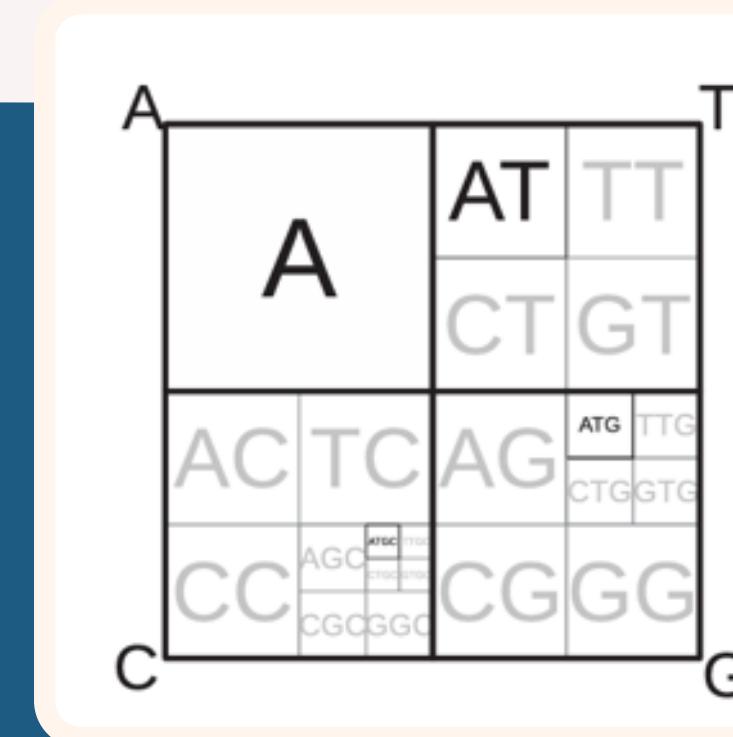
Si utilizza un CGR di forma quadrata , in cui i quattro angoli hanno il nome di ciascuna base.

01



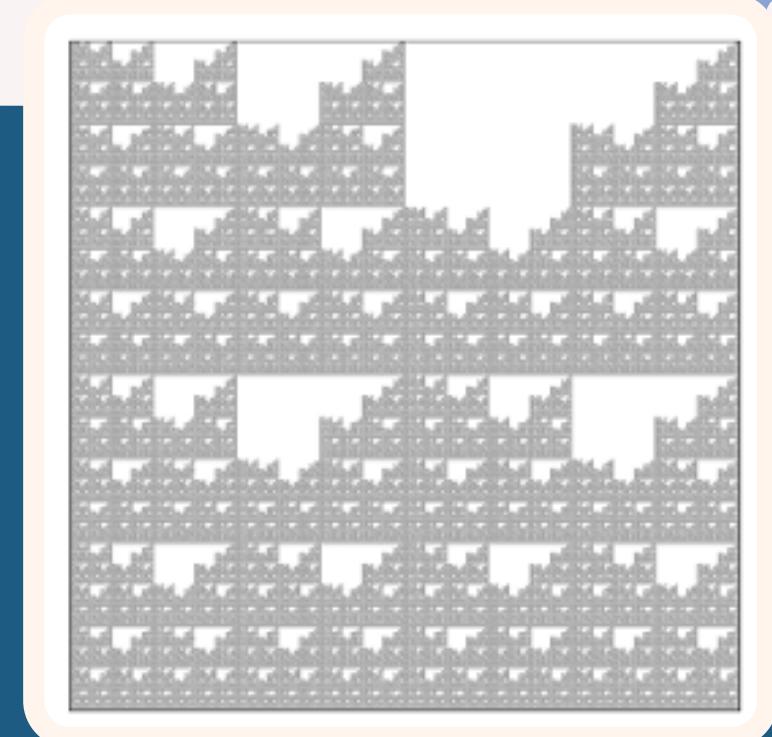
Vengono assegnati i nucleotidi alle coordinate CGR come segue: A è assegnato a (-1,1), T è assegnato a (1,1), C è assegnato a (-1,-1) e G è assegnato a (1, -1).

02



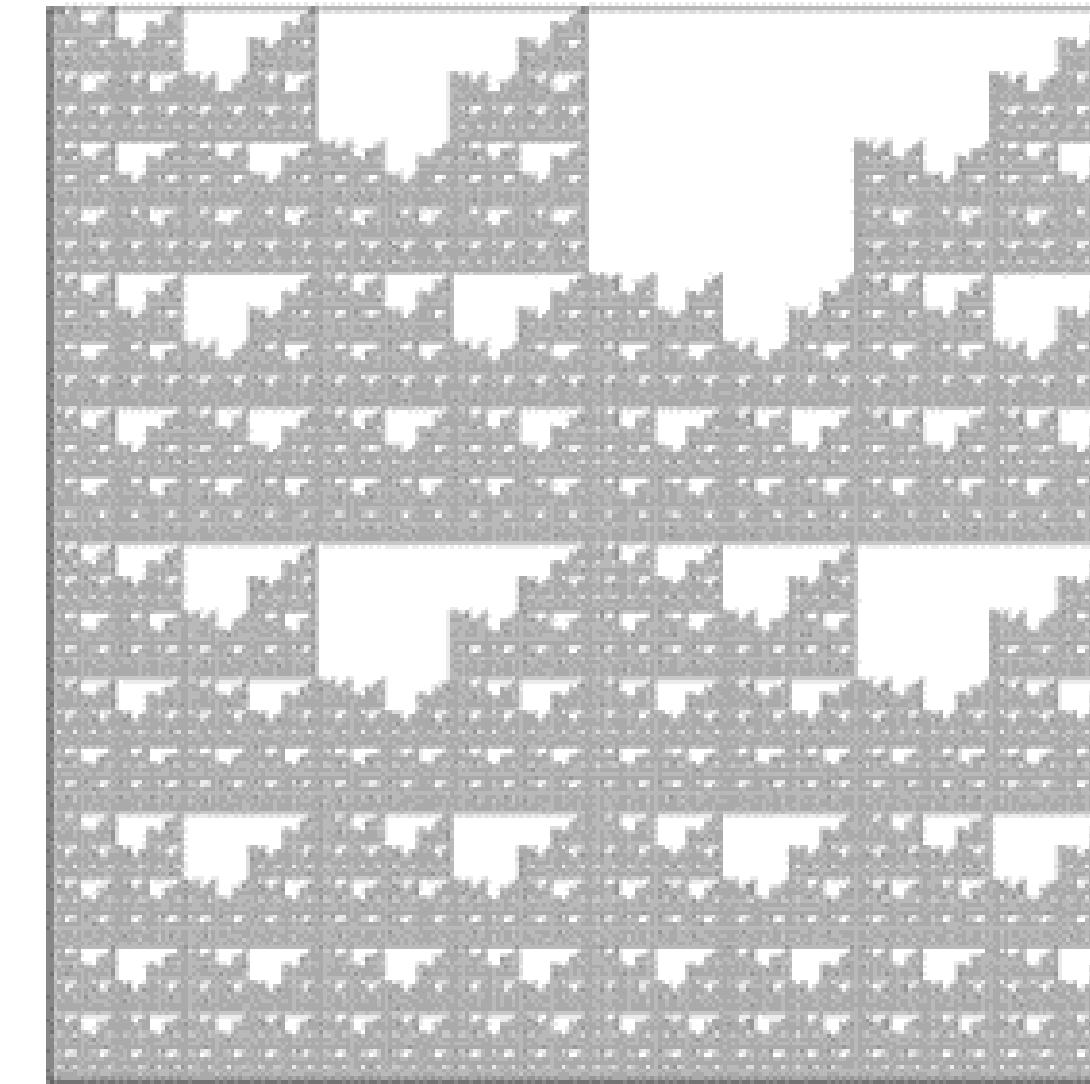
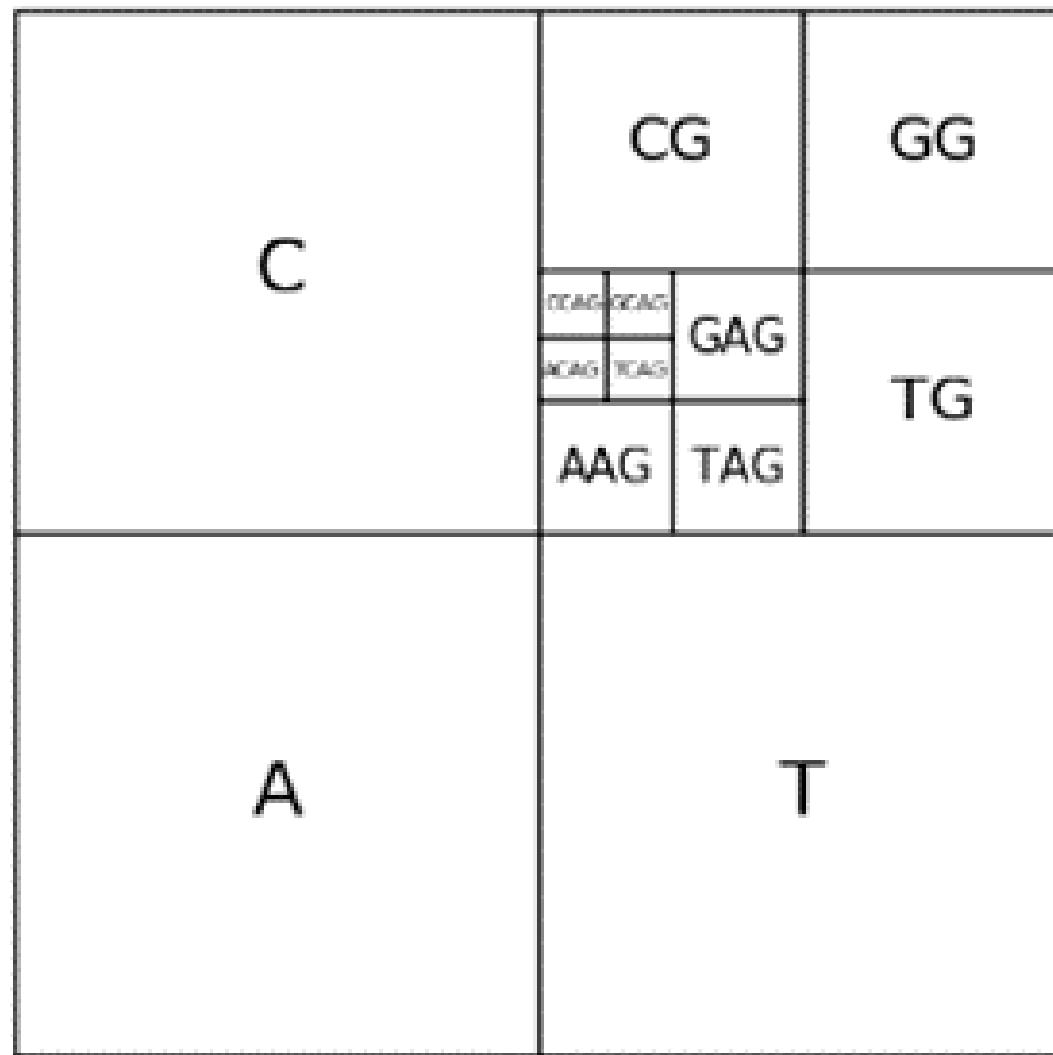
Secondo Jeffrey, ogni punto della rappresentazione CGR corrisponde esattamente a una sottosequenza (a partire dalla prima base), fino alla risoluzione dello schermo.

03



Bisogna tenere presente che qualsiasi base verrà sempre tracciata da qualche parte nel quadrante con la sua etichetta, poiché una base viene sempre tracciata a metà verso il suo angolo.

CGR AL DNA (3)



Il pattern più prominente è definito “double scoop”, appare in quasi tutte le sequenze DNA di vertebrati. Questo pattern è dovuto al fatto che c'è una relativa scarsità di guanina a seguito della citosina nella sequenza genica poiché i dinucleotidi CG sono inclini alla metilazione e successivamente alla mutazione.

CGR AL DNA(4)

Caratteristiche CGR di una sequenza DNA

Unicità

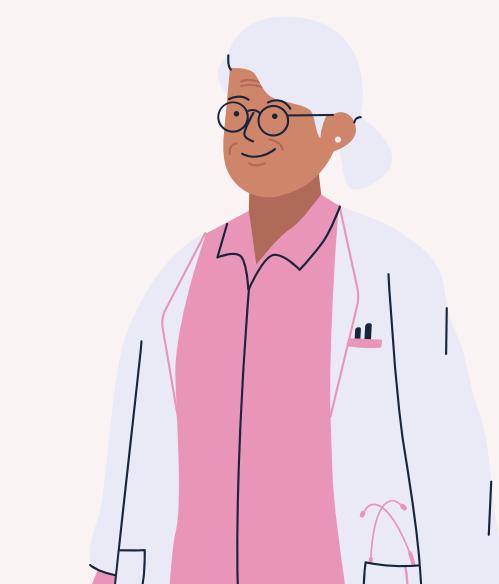
- Il punto k-esimo tracciato sul CGR di una sequenza corrisponde alla prima sottosequenza iniziale della sequenza lunga k, e a nessun'altra successiva (fino alla risoluzione dello schermo). Quindi, è presente una corrispondenza biunivoca tra le sottosequenze (ancorato all'inizio) di un gene e punti del CGR.

Limiti risolutivi

- La risoluzione dello schermo limita i dettagli. Tuttavia, qualsiasi parte dell'immagine può essere ingrandita, rivelando una struttura più fine.
- Questo ingrandimento è illimitato (purché ci siano più basi nella sequenza)

Vicinanza

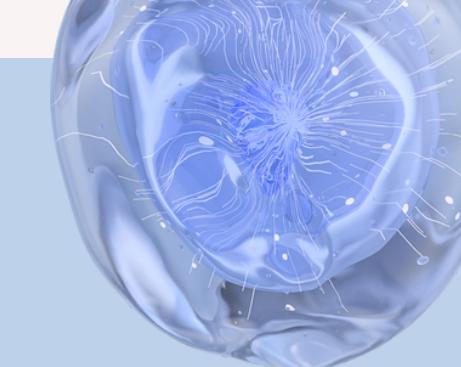
- Basi adiacenti nella sequenza non vengono plottate in maniera adiacente l'un l'altra ;
- Per cui, essere vicini nella rappresentazione CGR non significa essere vicini nella sequenza. La distanza euclidea nella CGR implica una nuova metrica o sottosequenza, o basi.





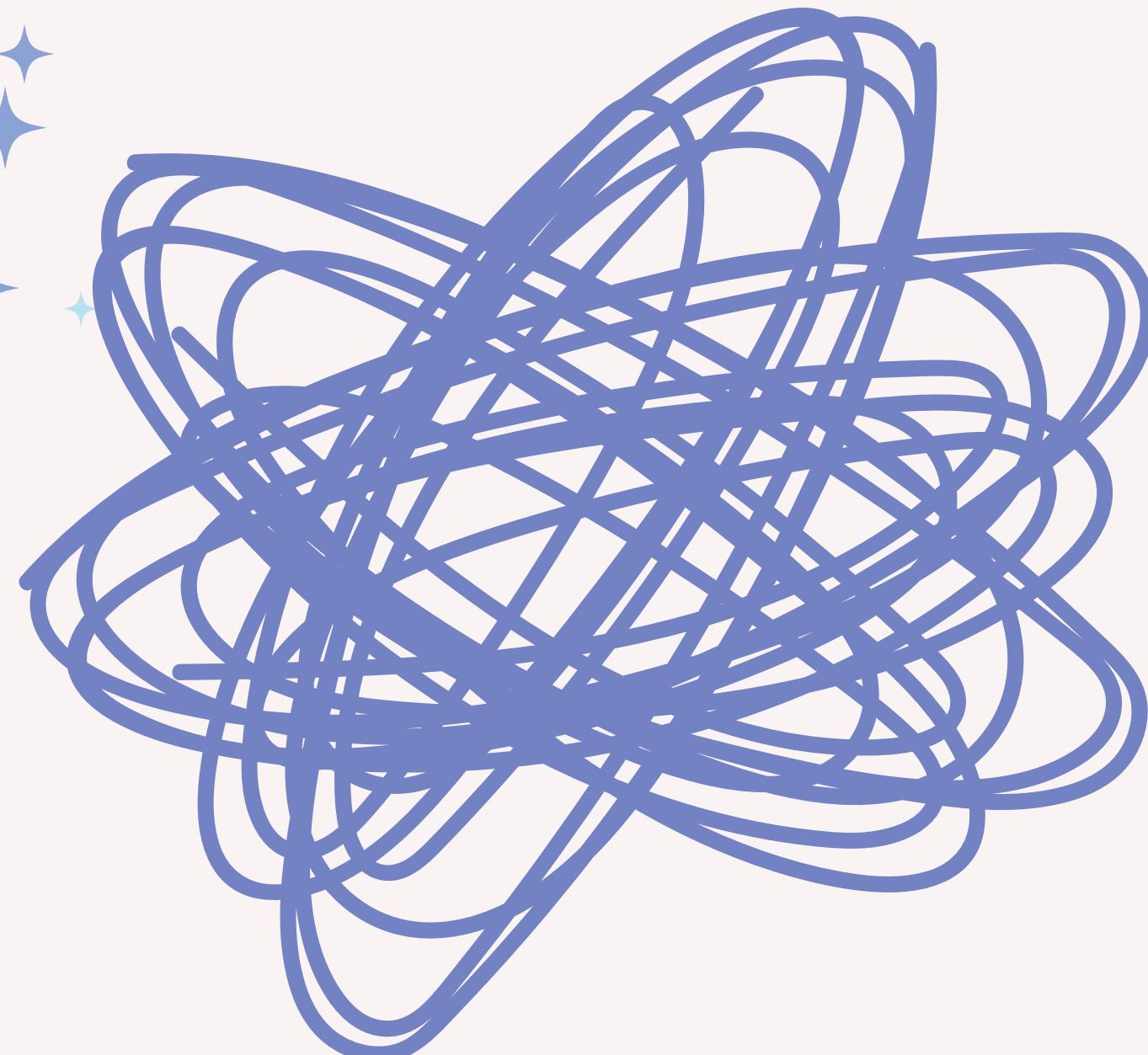
2. Frequency Chaos Game Representation

La Frequency Chaos Game Representation (FCGR) è un'estensione del metodo Chaos Game Representation (CGR) utilizzato per rappresentare graficamente sequenze di dati, come quelle genetiche. Mentre il CGR rappresenta una sequenza di simboli come un pattern frattale, la FCGR introduce un ulteriore livello di informazione considerando la frequenza con cui appaiono determinati sottosequenze all'interno della sequenza originale.



FCGR

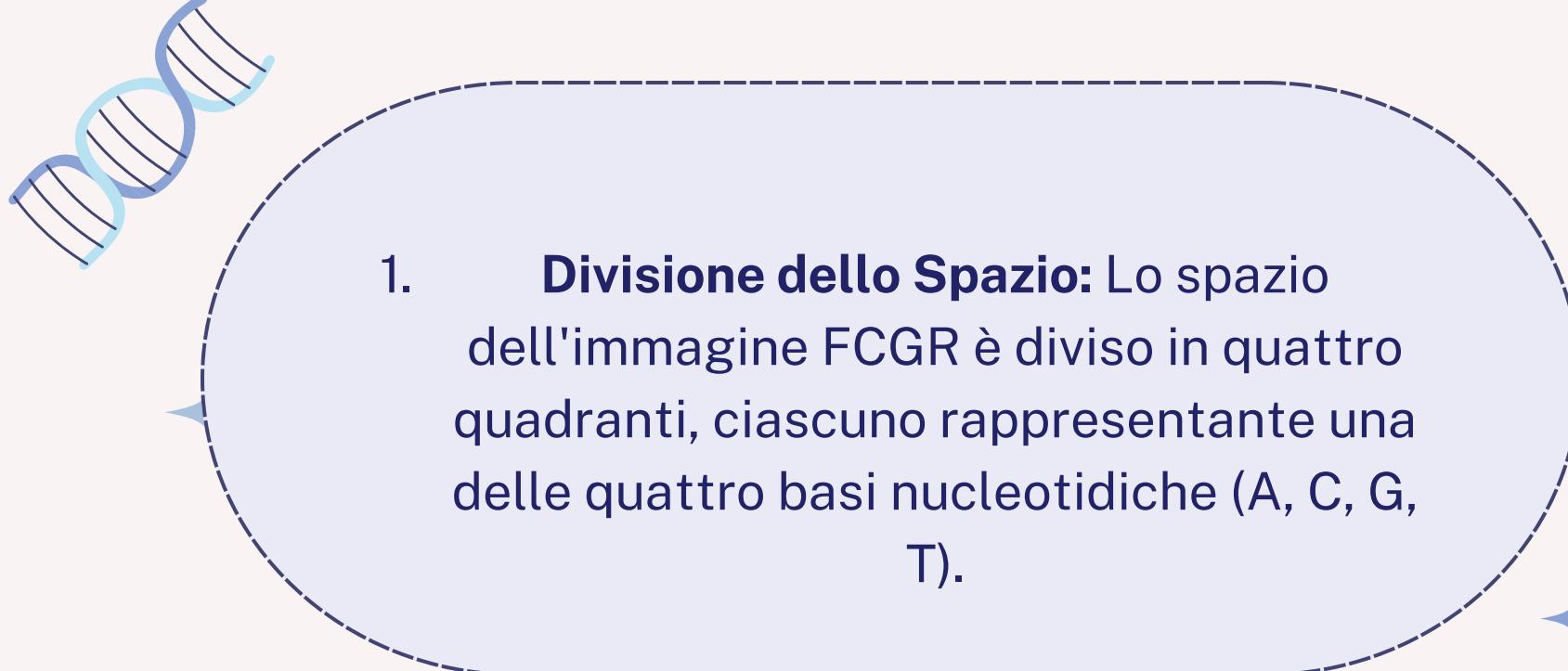
Frequency Chaos Game Representation



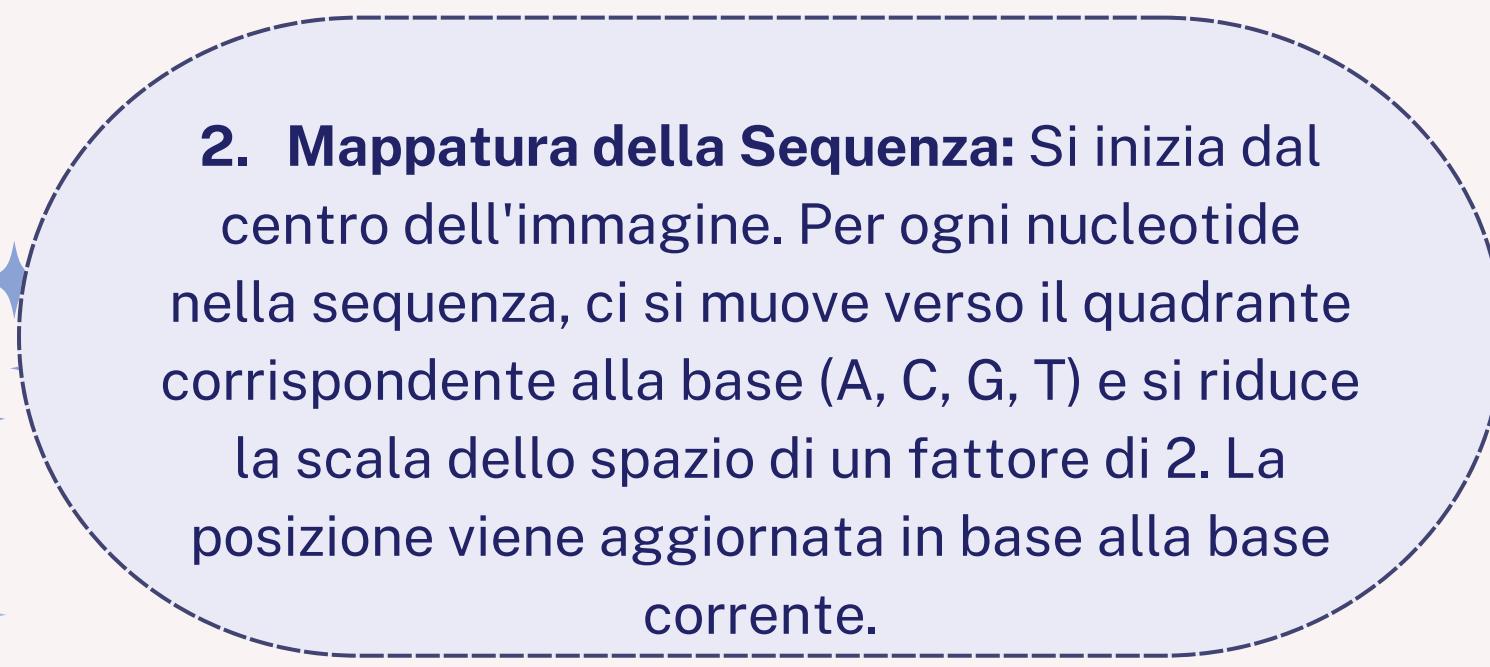
Mentre il CGR originale utilizza le coordinate esatte per ciascun punto, la discretizzazione viene chiamata FCGR. La FCGR si basa sul conteggio dei punti della CGR tenendo conto di una griglia predefinita. Questa procedura dà come risultato una matrice che rappresenta la frequenza di k-mers , e quindi una visualizzazione, sulla scala di grigi.

Costruzione di una

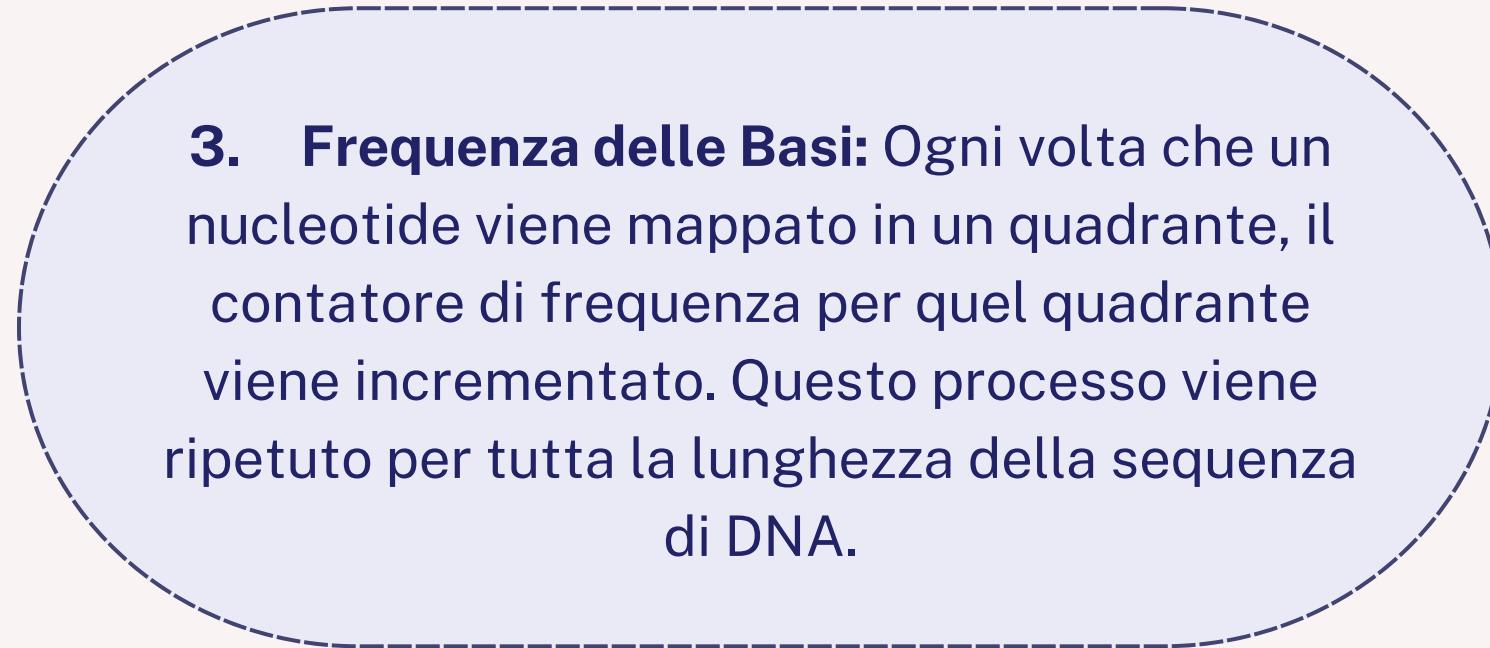
FCGR



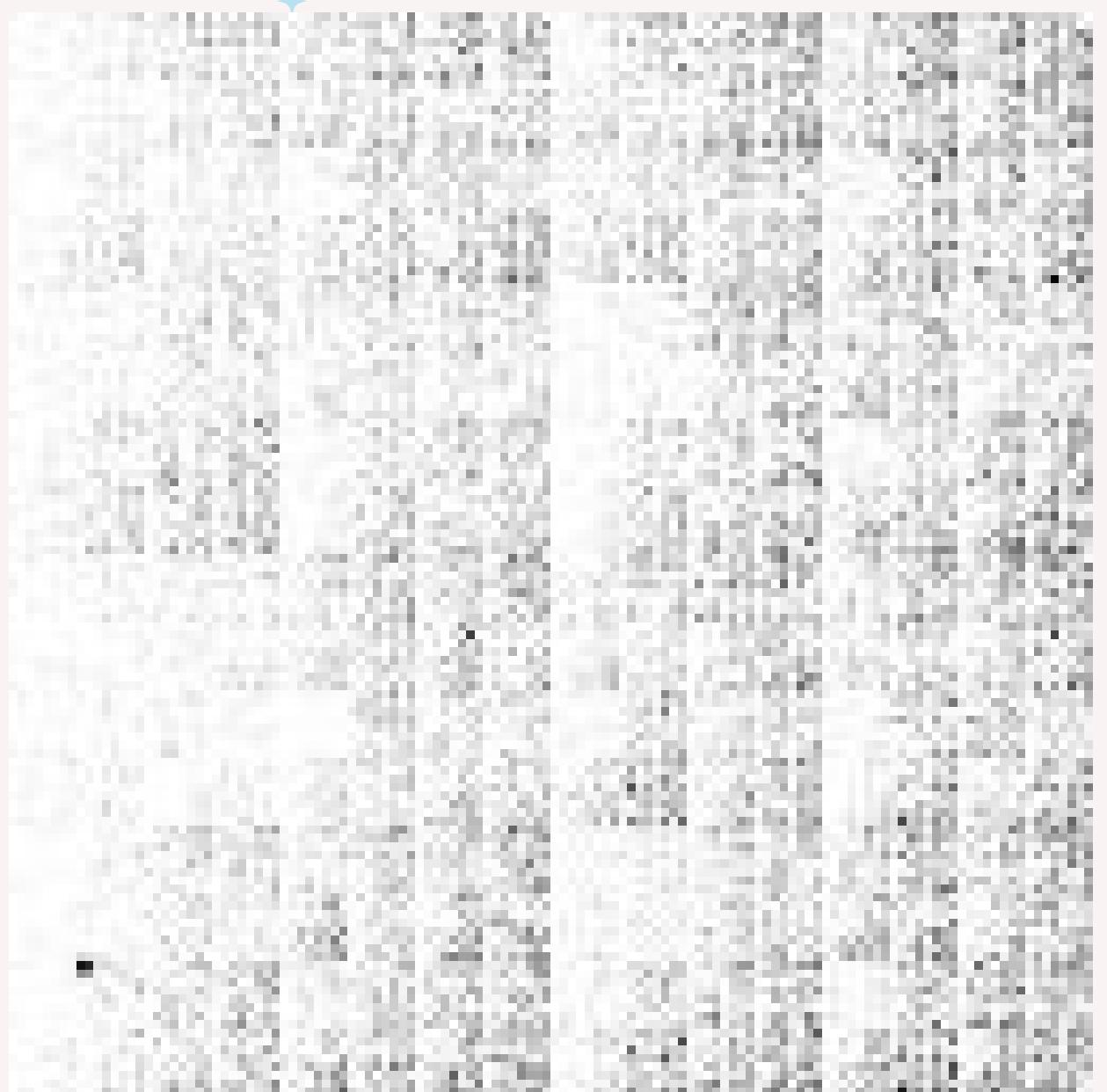
1. **Divisione dello Spazio:** Lo spazio dell'immagine FCGR è diviso in quattro quadranti, ciascuno rappresentante una delle quattro basi nucleotidiche (A, C, G, T).



2. **Mappatura della Sequenza:** Si inizia dal centro dell'immagine. Per ogni nucleotide nella sequenza, ci si muove verso il quadrante corrispondente alla base (A, C, G, T) e si riduce la scala dello spazio di un fattore di 2. La posizione viene aggiornata in base alla base corrente.



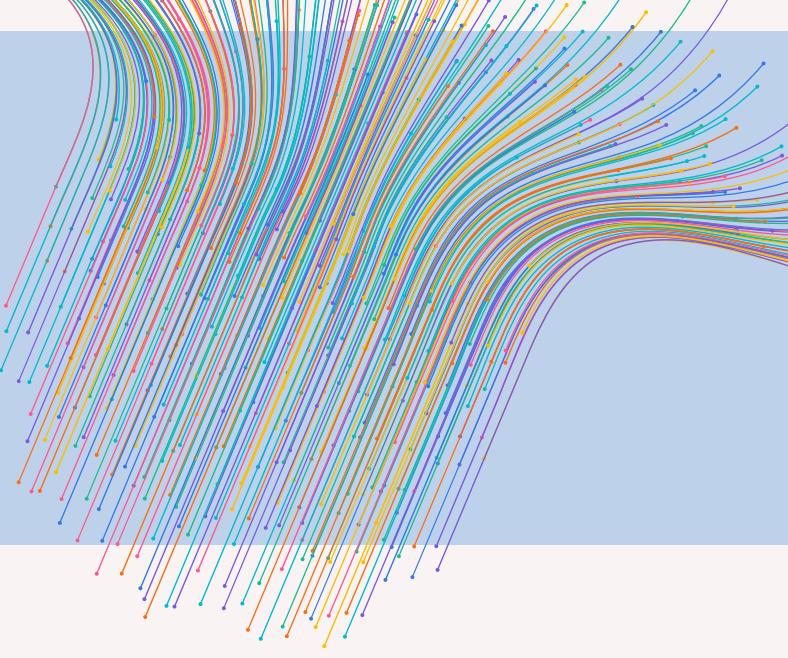
3. **Frequenza delle Basi:** Ogni volta che un nucleotide viene mappato in un quadrante, il contatore di frequenza per quel quadrante viene incrementato. Questo processo viene ripetuto per tutta la lunghezza della sequenza di DNA.





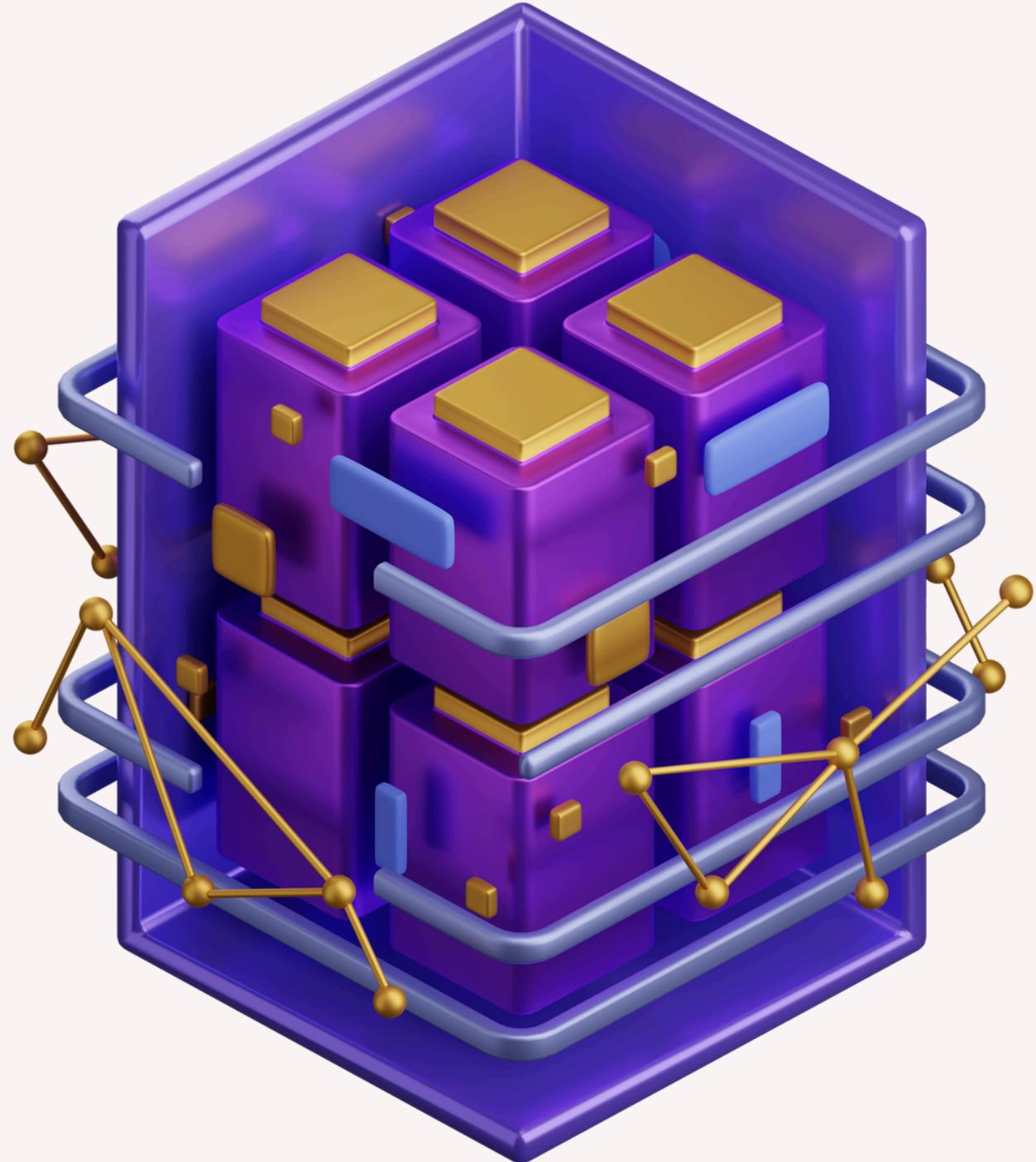
3. Rete Neurale Convoluzionale

Le reti neurali convoluzionali (CNN) sono un tipo di rete neurale artificiale progettata per elaborare e analizzare dati strutturati in griglie, come immagini. Sono particolarmente efficaci per compiti di visione artificiale, come il riconoscimento di oggetti, la classificazione di immagini e il rilevamento di volti.



CNN

Funzionamento



Strato convoluzionale

Il cuore di una CNN è il processo di convoluzione, dove un filtro (o kernel) scorre sull'immagine di input. Ogni filtro è una matrice di pesi che "estrai" caratteristiche specifiche, come bordi, angoli o texture.

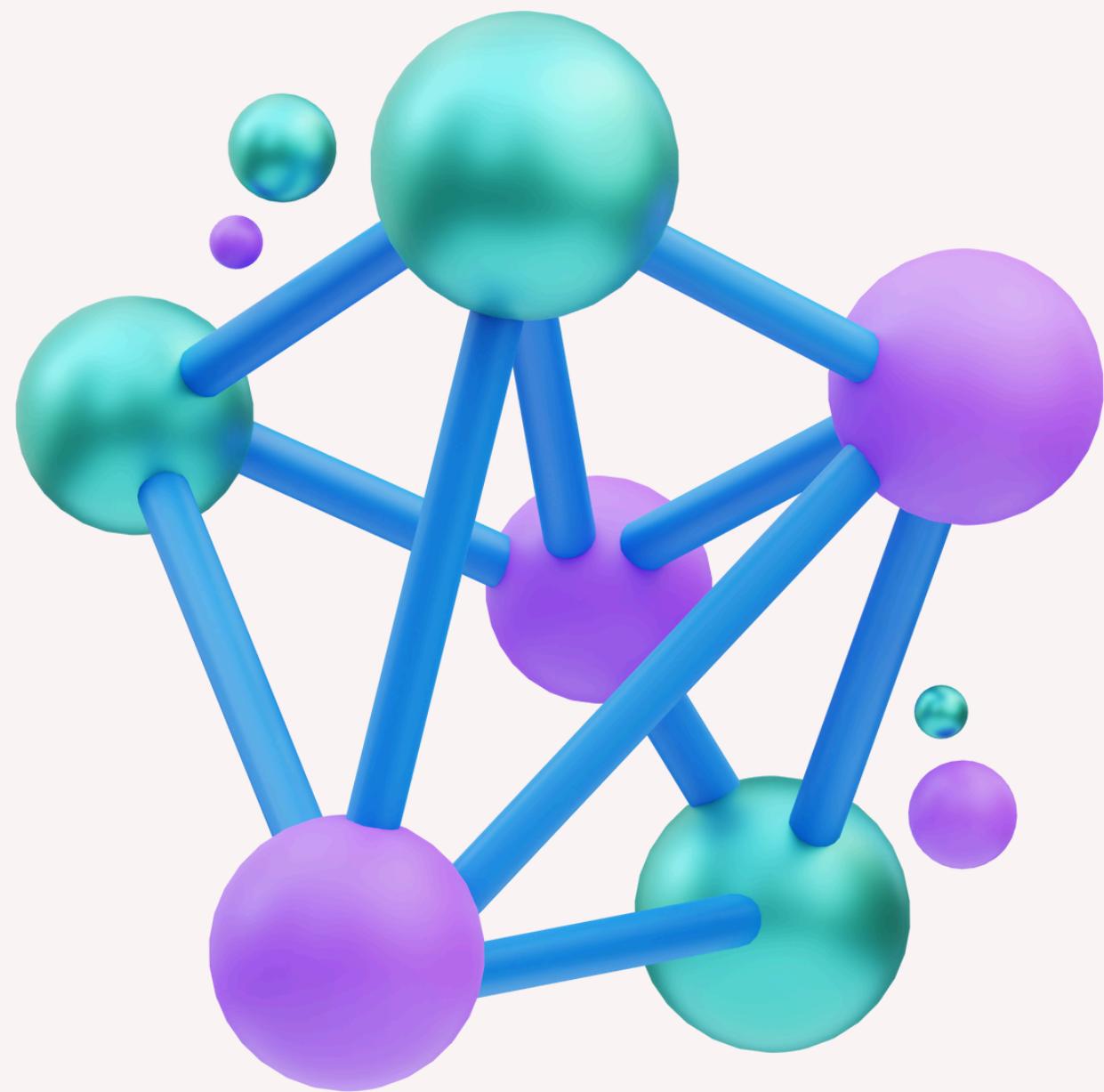
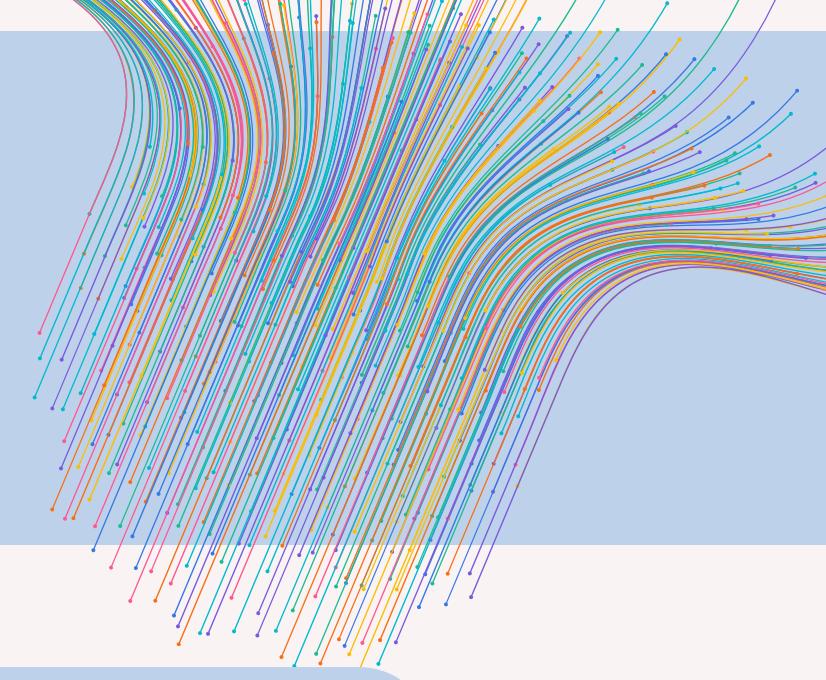
Strato di attivazione

Dopo la convoluzione, viene applicata una funzione di attivazione ReLU, che introduce non linearità nel modello. ReLU sostituisce i valori negativi nella mappa delle caratteristiche con zero, mantenendo solo le attivazioni positive.



CNN

Funzionamento(2)

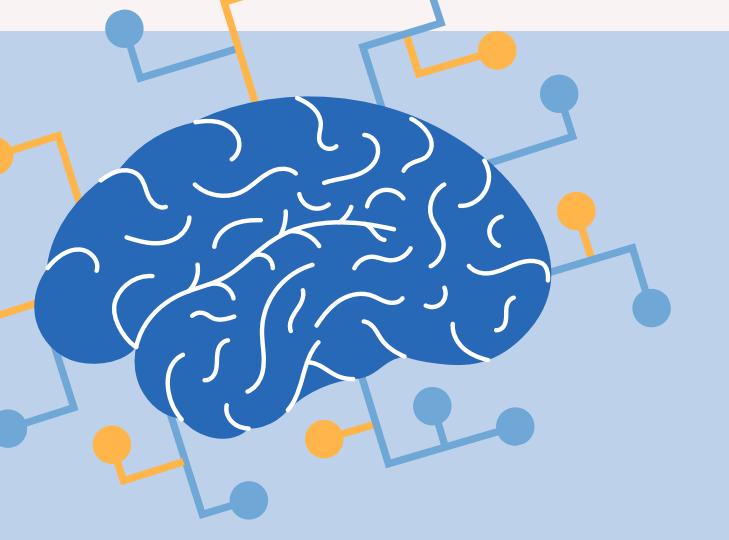


Strato pooling

Questo strato riduce la risoluzione dell'immagine, prendendo il valore massimo all'interno di piccoli blocchi della mappa delle caratteristiche. Il pooling mantiene le caratteristiche più importanti riducendo il numero di parametri e la complessità computazionale.

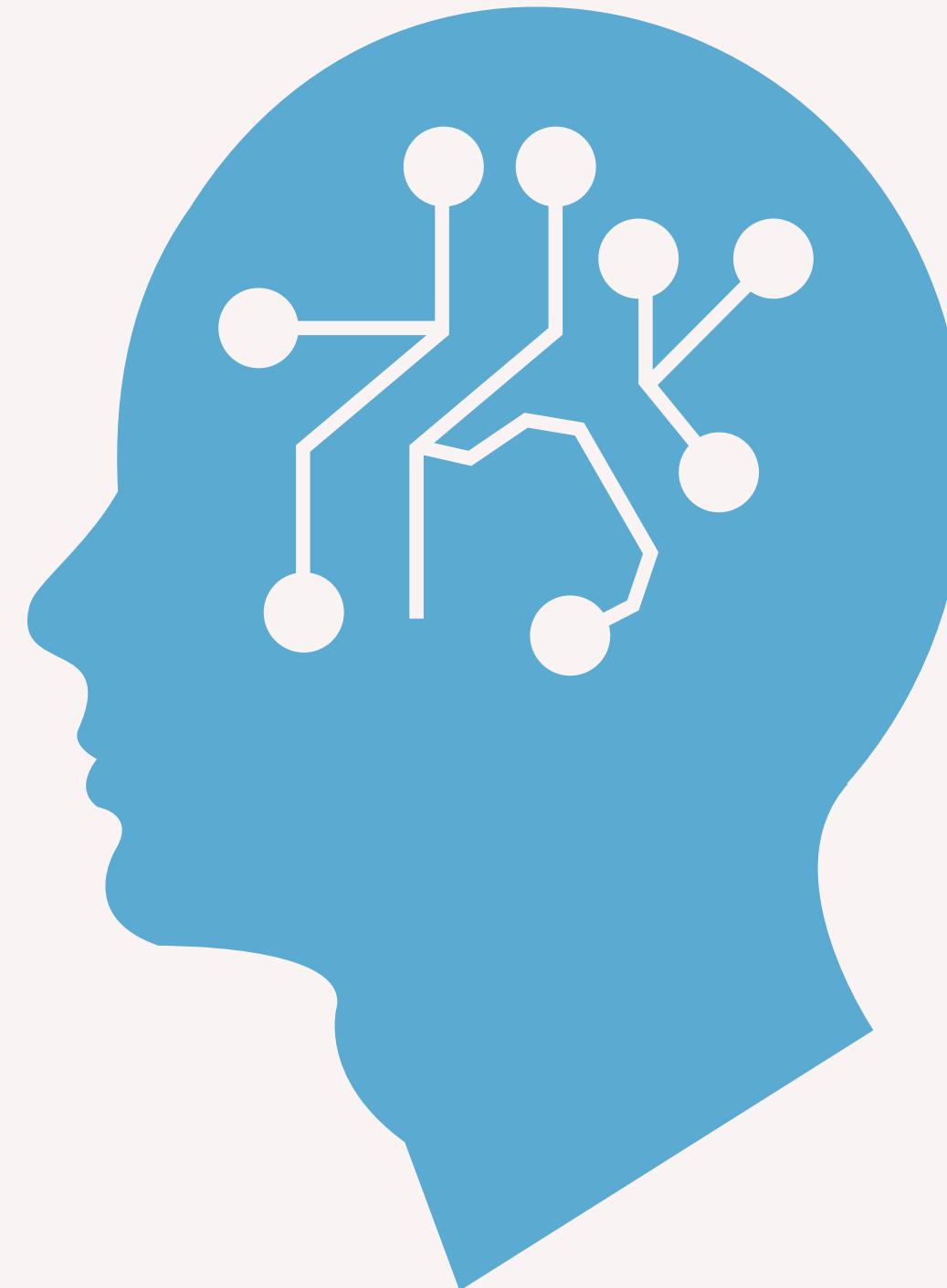
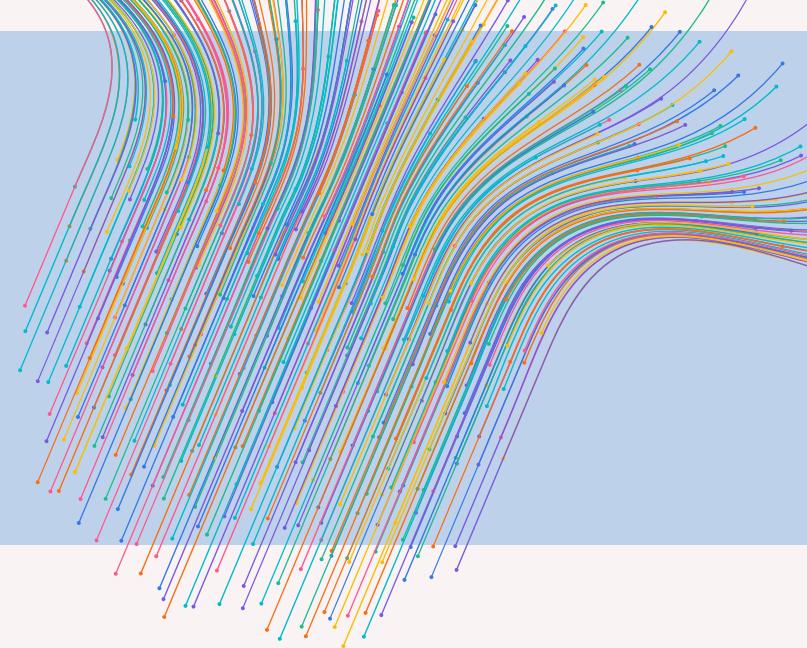
Classificazione

la CNN appiattisce la mappa delle caratteristiche in un vettore. Questo vettore viene quindi passato attraverso strati completamente connessi (come in una rete neurale tradizionale), dove ogni neurone è connesso a tutti i neuroni dello strato precedente.



CNN

Funzionamento(3)



Backpropagation

Questo avviene confrontando l'output predetto con l'etichetta corretta e calcolando l'errore, che viene poi propagato all'indietro nella rete per correggere i pesi.

Output

Il risultato finale della CNN è una serie di probabilità che indicano la classe più probabile dell'immagine di input.

03. PRESENTAZIONE DEI RISULTATI





1. DATASET



- I dati relativi alle sequenze genomiche e ai metadati associati sono stati resi disponibili pubblicamente attraverso il database EpiCov di GISAID.
- Pazienti sequenziati nel periodo compreso tra il 24 dicembre 2019 e il 28 febbraio 2022.
- Le sequenze sono state raccolte da un'ampia gamma di paesi e territori, totalizzando contributi da circa 211 diverse nazioni e regioni del mondo.



Risultati rilevanti dal test del chi-quadrato

Genere

In totale sono stati osservati 369181 pazienti, di cui il 52.7% sono maschi e 47.3% sono femmine. La distribuzione della rilevazione della variante prevede una prevalenza di uomini rispetto alle donne.

F
47.3%



M
52.7%

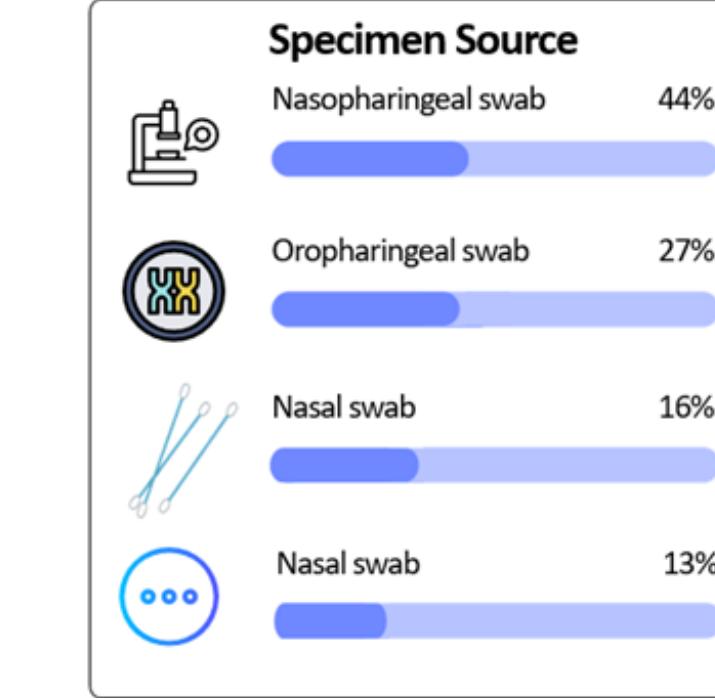
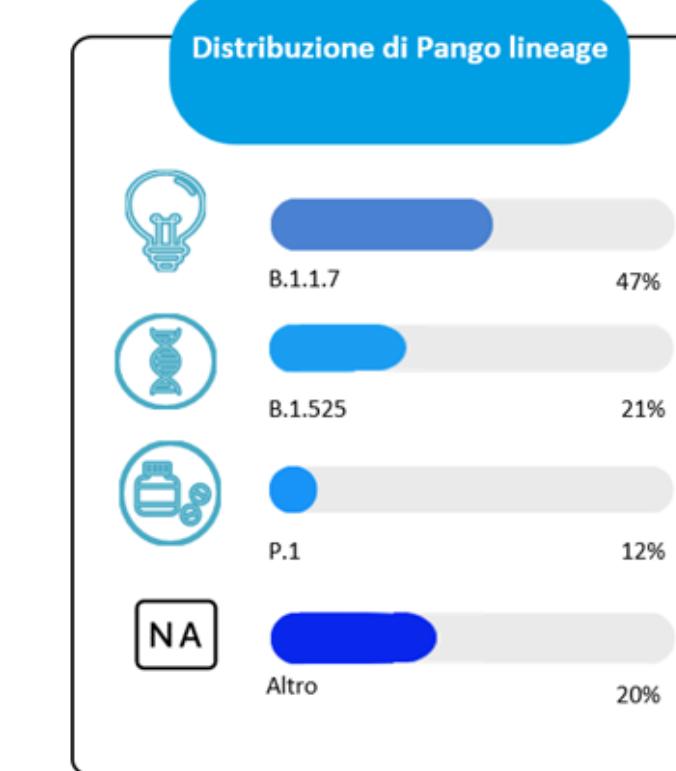
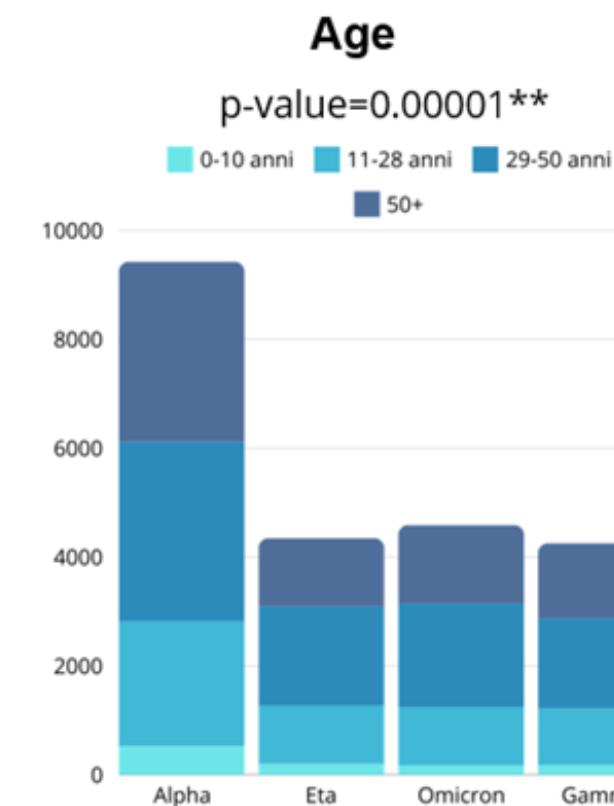
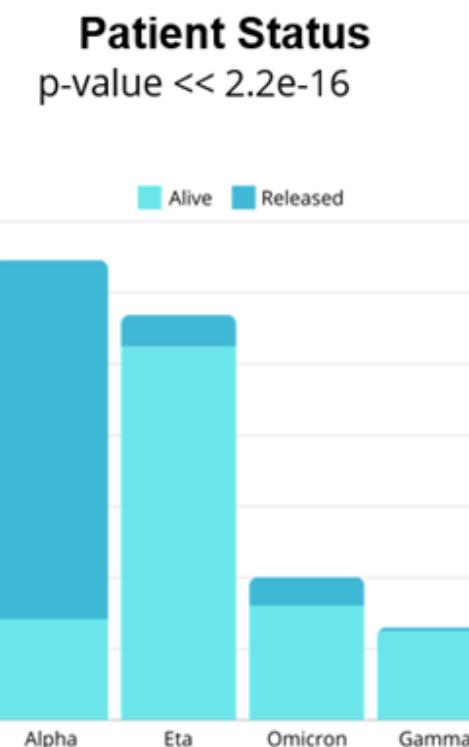
Classe O e S

La prevalenza di clade O nell'intera coorte è del 71%

71% 29%



Risultati rilevanti dal test del chi-quadrato(2)



2. Analisi delle mutazioni

EDA 1

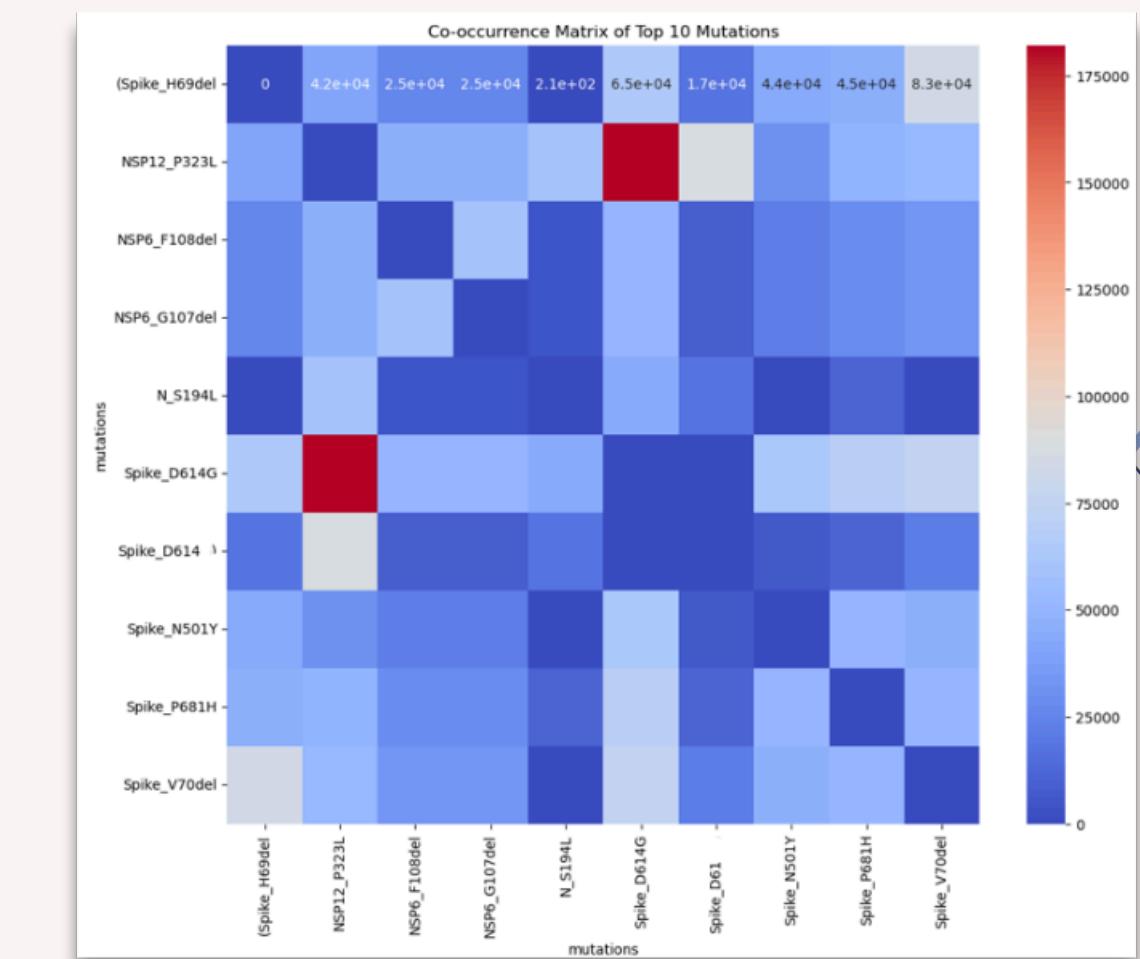
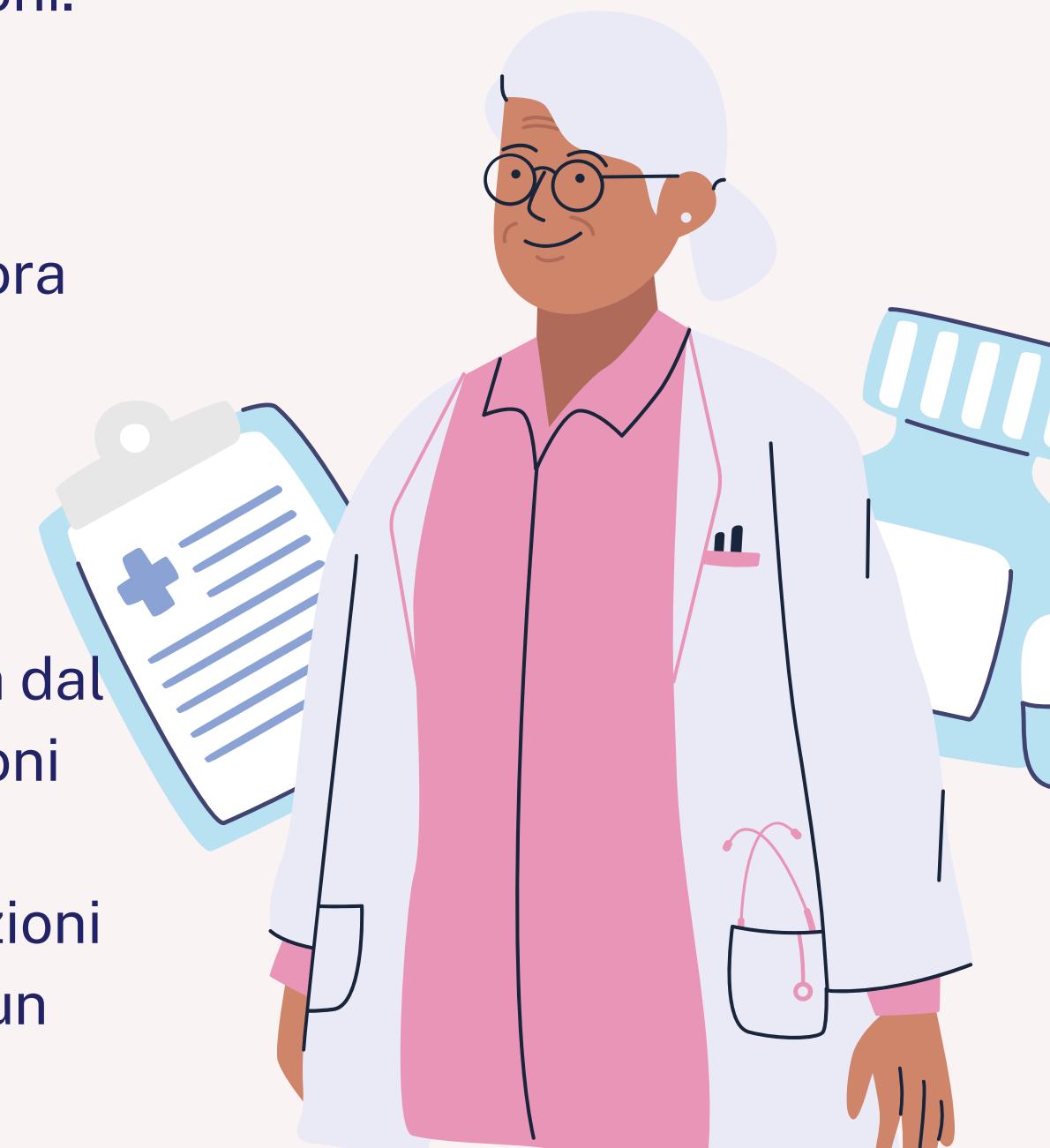
Alcune proteine, come le proteine NSP3 e NSP12, presentano un numero relativamente elevato di mutazioni, con Spike che mostra il conteggio più alto tra tutti con oltre 1.75×10^6 mutazioni.

EDA 2

La mutazione NSP12_P323L è la più comune, seguita da due versioni della mutazione Spike_D614G. Altre mutazioni frequenti includono Spike_V70del e Spike_H69del, che sono associati alla proteina Spike.

EDA 3

La matrice di co-occurenza esplora quanto spesso due mutazioni specifiche si presentano simultaneamente nello stesso campione. E' presente una co-occorrenza molto elevata (indicata dal colore rosso scuro) tra le mutazioni Spike_D614G e NSP12_P323L, suggerendo che queste due mutazioni tendono a verificarsi insieme in un gran numero di campioni.





La proteina Spike

La proteina spike di SARS-CoV-2 è il principale meccanismo che il virus utilizza per infettare le cellule bersaglio; questa proteina è formata da due componenti principali: la subunità S1 e la subunità S2.





La proteina Spike

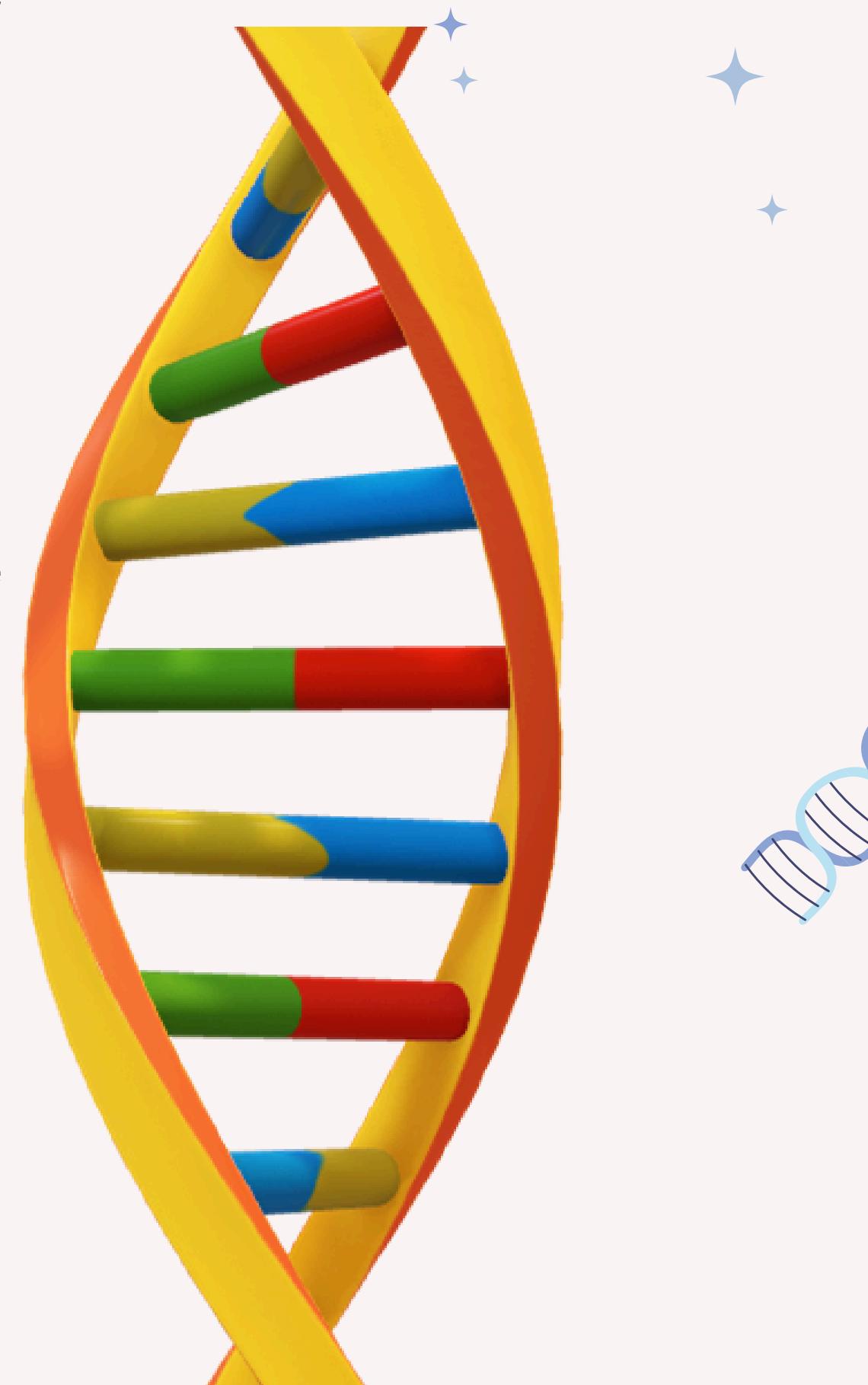
La proteina spike di SARS-CoV-2 è il principale meccanismo che il virus utilizza per infettare le cellule bersaglio; questa proteina è formata da due componenti principali: la subunità S1 e la subunità S2.

La subunità S1 è una regione molto flessibile e contiene il meccanismo chiamato RBD, attraverso il quale il virus è in grado di riconoscere e legare il recettore ACE2, che è la porta di ingresso del virus nelle cellule del nostro organismo.

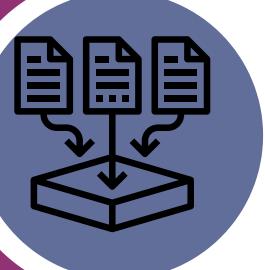
La subunità S2 contiene una piccola regione chiamata FP, che è "l'ago" attraverso il quale il virus riesce a penetrare nella cellula bersaglio.

Una volta che la subunità S1 della proteina spike ha legato il recettore, la subunità S2 cambia forma e "conficca" la regione FP nella membrana della cellula ospite, dando inizio al processo di invasione.

Per via della sua fondamentale importanza nel processo di infezione, la proteina spike di SARS-CoV-2 è uno dei bersagli farmacologici più studiati.



3. ARCHITETTURA E PROBLEMA



Vasta Quantità di dati generati



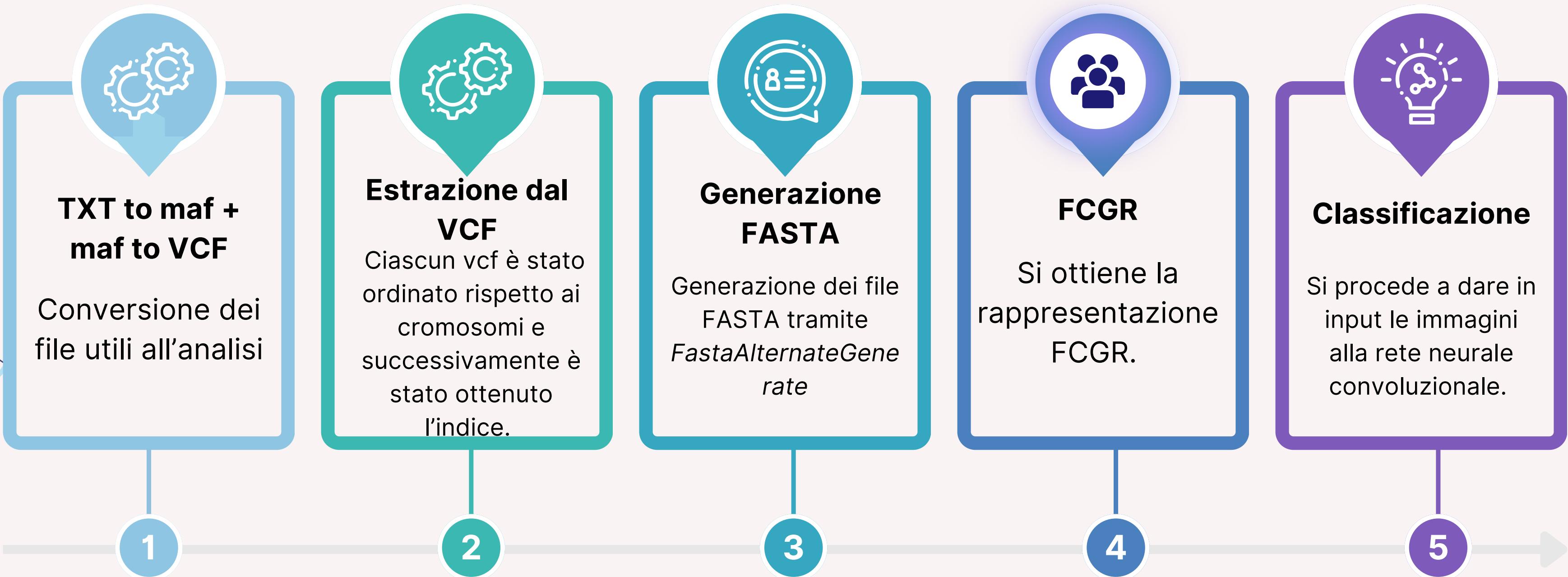
Manutenzione e conservazione



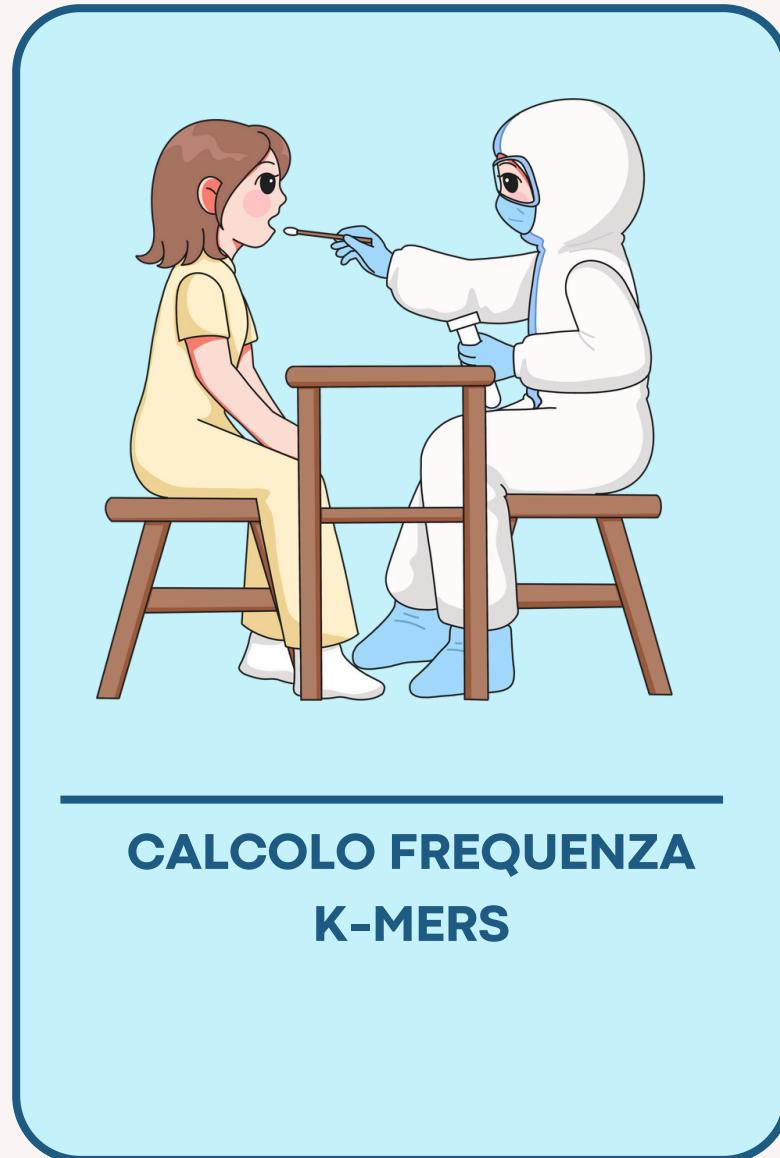
Privacy e etica



PROTOCOLLO Sperimentale



1. GENERAZIONE FCGR



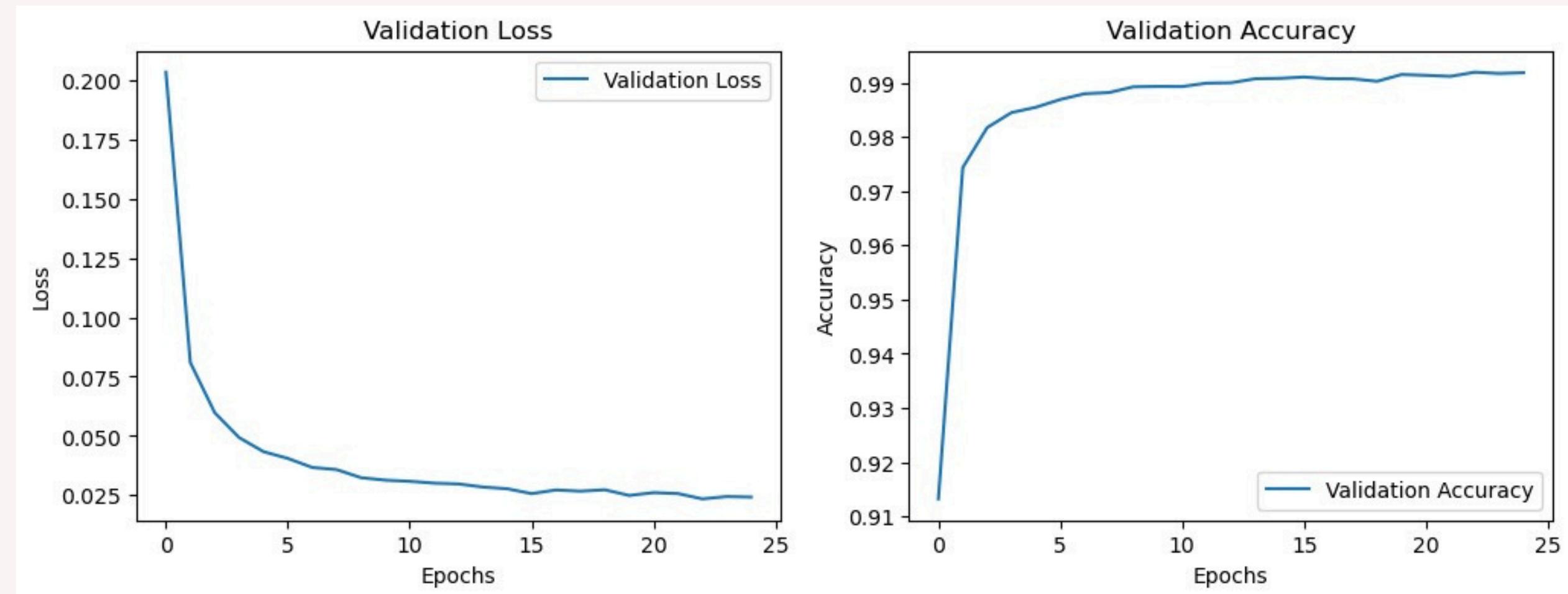
2. COSTRUZIONE CNN

PARAMETRI SCELTI

- Ogni conv2D utilizza 4 filtri e un kernel di dimensioni 8x8 pixel;
- Dopo ogni strato convoluzionale è stata impiegata la funzione di attivazione ReLU;
- ciascun strato convoluzionale è seguito da un'operazione di normalizzazione, che standardizza l'output dello strato convoluzionale;
- ogni strato convoluzionale era seguito da un'operazione di Max Pooling con finestra 2x2.



3. RISULTATI CNN



3. RISULTATI CNN



Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 1.00 | 0.99 | 0.99 | 7672 |
| 1 | 0.98 | 0.99 | 0.98 | 3115 |
| accuracy | | | 0.99 | 10787 |
| macro avg | 0.99 | 0.99 | 0.99 | 10787 |
| weighted avg | 0.99 | 0.99 | 0.99 | 10787 |

Confusion Matrix:

```
[[7596 76]
 [ 30 3085]]
```

Accuracy: 0.9901733568183925

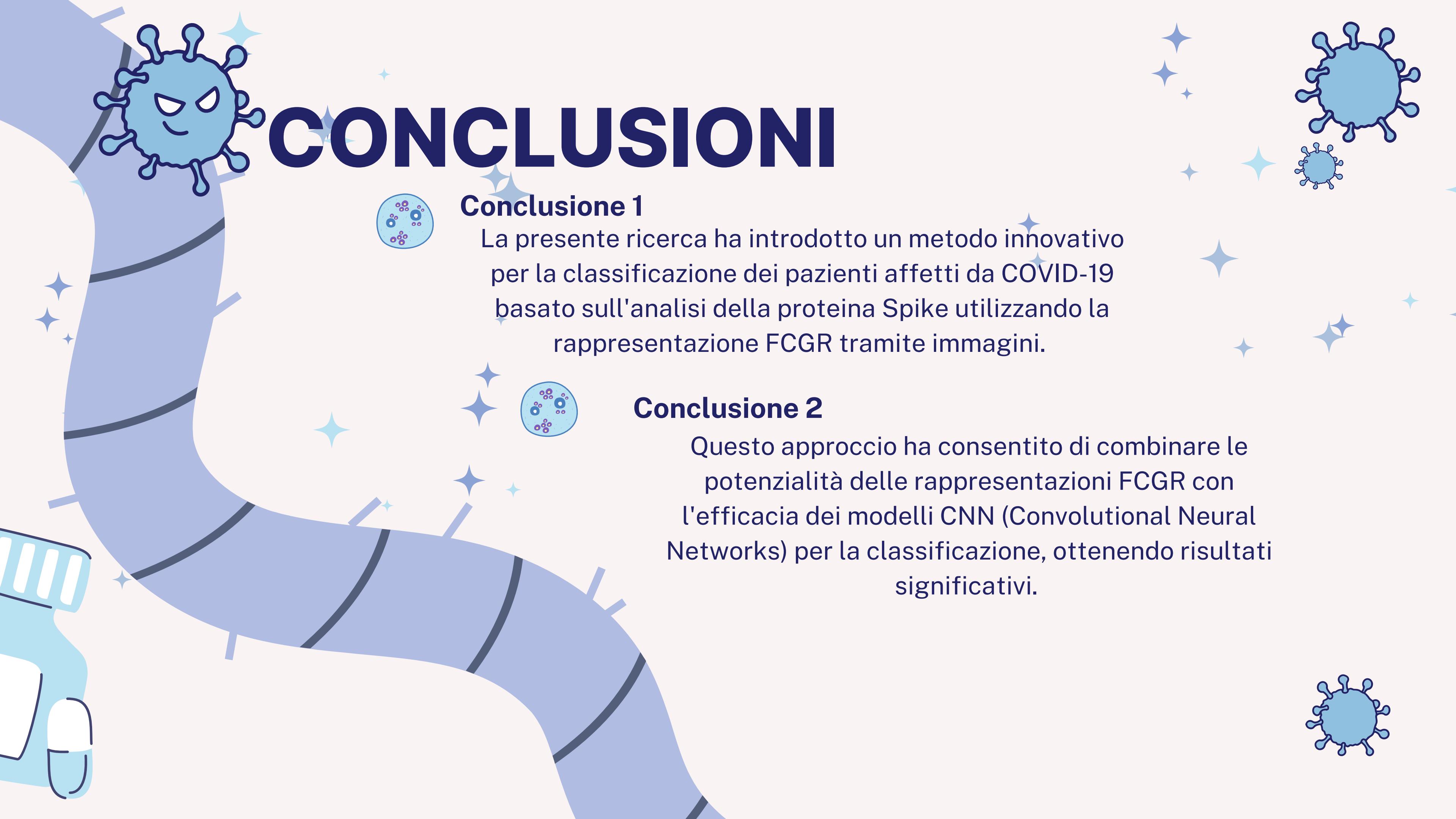
F1 Score: 0.9901945950451706

Precision: 0.9902591099700025

Recall: 0.9901733568183925

04. CONCLUSIONI E SVILUPPI FUTURI





CONCLUSIONI

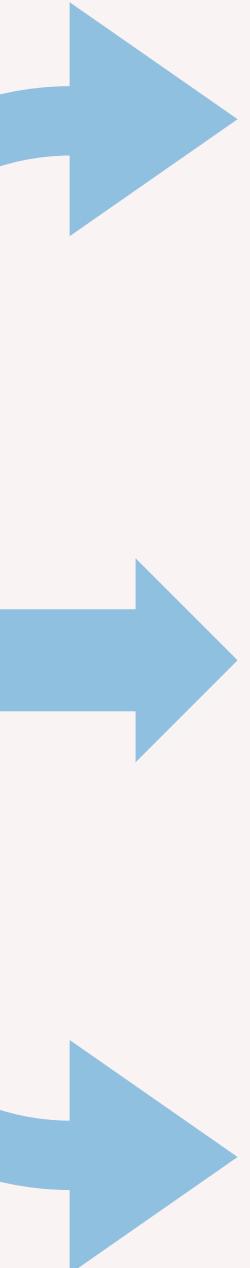
Conclusione 1

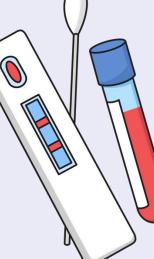
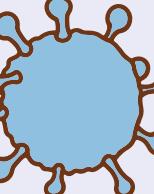
La presente ricerca ha introdotto un metodo innovativo per la classificazione dei pazienti affetti da COVID-19 basato sull'analisi della proteina Spike utilizzando la rappresentazione FCGR tramite immagini.

Conclusione 2

Questo approccio ha consentito di combinare le potenzialità delle rappresentazioni FCGR con l'efficacia dei modelli CNN (Convolutional Neural Networks) per la classificazione, ottenendo risultati significativi.

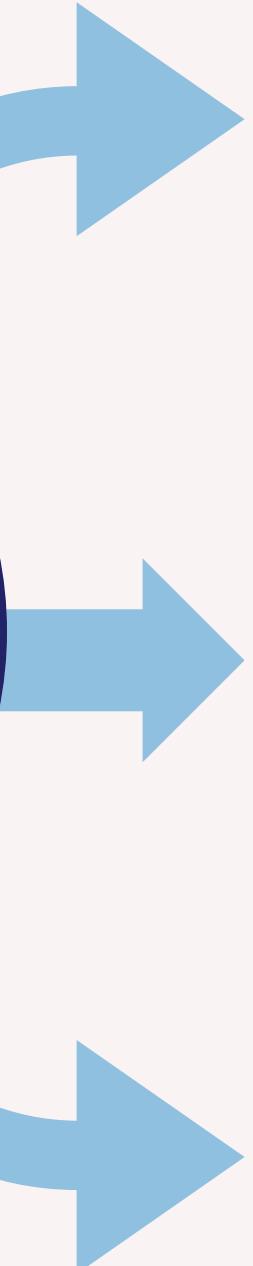
Sviluppi futuri

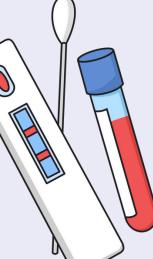
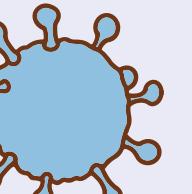


-  L'approccio basato su immagini FCGR potrebbe essere esteso per classificare non solo la proteina Spike, ma anche altre proteine virali.
-  L'identificazione di modelli CNN ottimizzati per rappresentazioni con un numero maggiore di k-mers potrebbe aumentare la risoluzione e la precisione della classificazione.
-  -L'esplorazione di architetture più avanzate, potrebbe migliorare ulteriormente le prestazioni del modello, specialmente nella gestione di immagini ad alta dimensionalità derivanti da rappresentazioni FCGR.



Sviluppi futuri (2)



-  Il modello potrebbe essere utilizzato per monitorare in tempo reale le mutazioni nelle sequenze proteiche, fornendo previsioni su come le mutazioni potrebbero influenzare la struttura e la funzione delle proteine.
-  Un altro sviluppo cruciale potrebbe essere l'espansione del dataset utilizzato per addestrare le reti neurali, includendo sequenze proteiche da una gamma più ampia di ceppi virali e varianti.
-  Lavorare in collaborazione con istituti di ricerca a livello globale potrebbe fornire una base di dati più ricca e diversificata, migliorando ulteriormente la capacità del modello di generalizzare e fornire previsioni affidabili.



Grazie per l'attenzione

