



UNIVERSITÀ DEGLI STUDI DI SALERNO

Classificazione delle Sequenze della Proteina Spike del SARS-CoV-2 Tramite Rappresentazione FCGR e Reti Neurali Convoluzionali

Strumenti Formali per la Bioinformatica

Girolamo Martina - 0522501680
Nappi Severino - 0522501681
Ragozzini Emanuele - 0522502039

De Felice Clelia
Zaccagnino Rocco
Zizza Rosalba

Contenuti

01

INTRODUZIONE

02

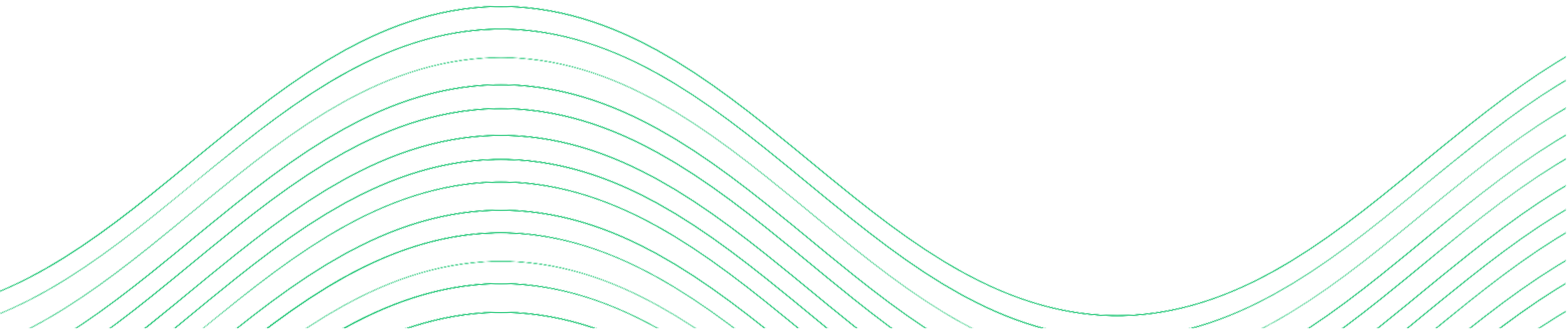
METODOLOGIA
UTILIZZATA

03

RISULTATI

04

CONCLUSIONI



Introduzione



Contesto: La pandemia da SARS-CoV-2 ha cambiato radicalmente la vita quotidiana dal 2020.

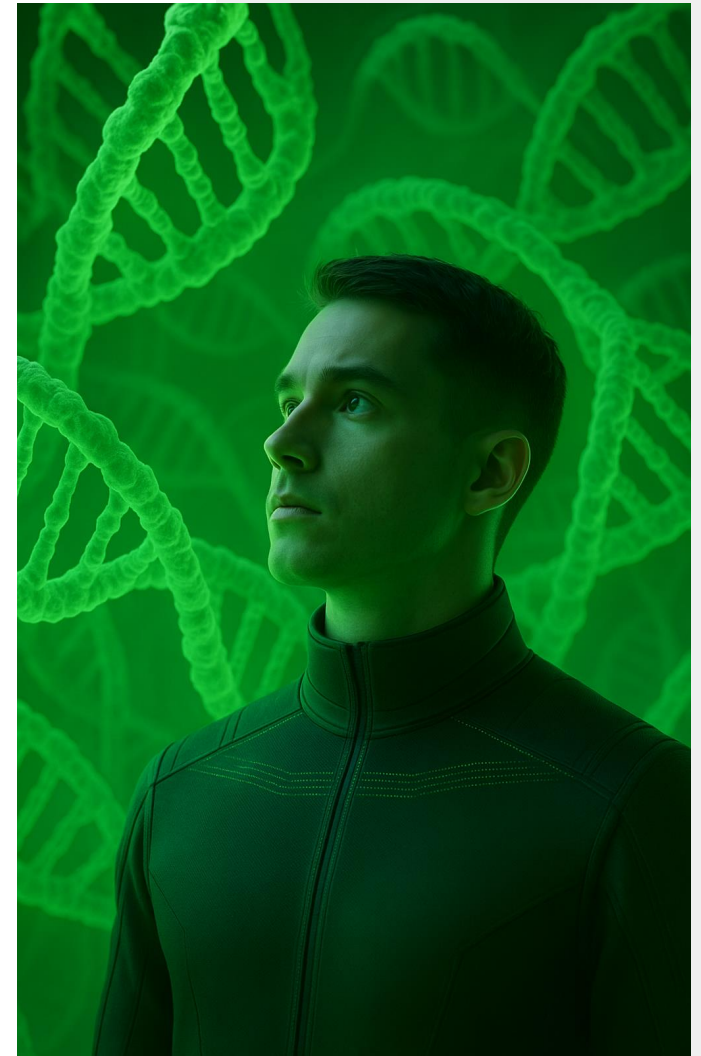


Obiettivo: Comprendere i meccanismi biologici e computazionali dell'evoluzione virale per migliorare prevenzione, diagnosi e terapia.



Approcci:

- Analisi molecolare (struttura e mutazioni genetiche)
- Approcci computazionali (modelli matematici, machine learning, interpretabilità dei modelli)



Introduzione

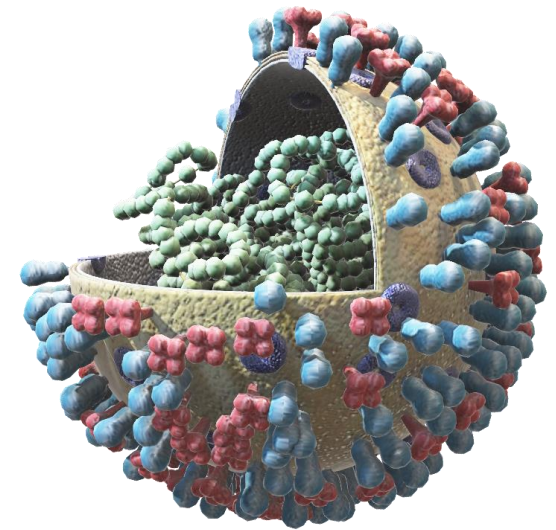
Virus: Caratteristiche Biologiche e Trasmissione

I virus sono costituiti da materiale genetico, rappresentato da DNA o RNA, racchiuso all'interno di un capsido proteico. In alcune tipologie virali è inoltre presente un pericapside lipidico di natura facoltativa.

L'infezione virale inizia con il riconoscimento e il legame a specifici recettori espressi sulla superficie delle cellule ospiti, seguito dall'introduzione del materiale genetico virale all'interno della cellula. Il virus sfrutta successivamente i meccanismi replicativi dell'ospite per la produzione di nuove particelle virali.

I virus si classificano in base al tipo di acido nucleico. I virus a DNA tendono a mantenere una maggiore stabilità genetica, mentre i virus a RNA presentano un'elevata variabilità a causa della rapida mutazione. Gli RNA virali possono essere a singolo filamento positivo (+) o negativo (-), in base alla loro capacità di essere tradotti direttamente in proteine.

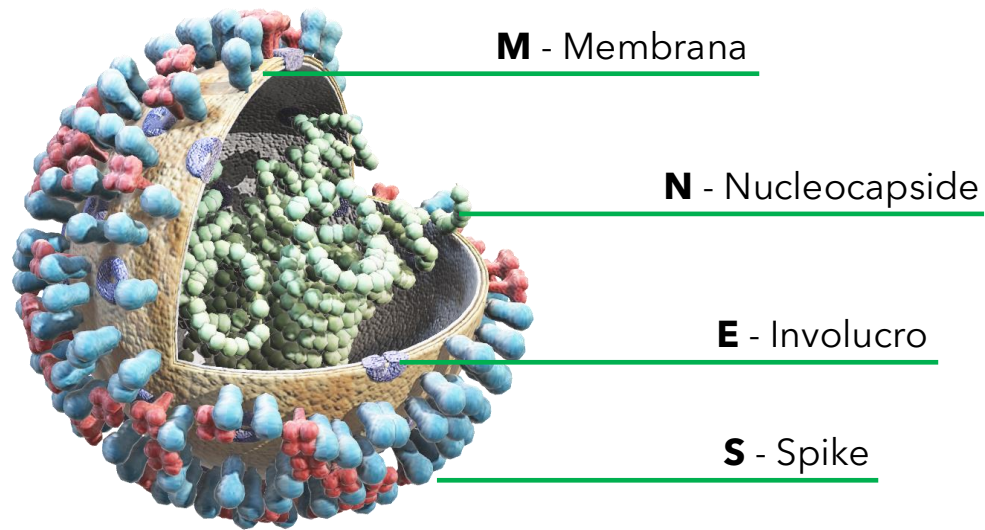
La trasmissione dei virus può avvenire per via aerea (ad esempio, virus influenzali), per via sessuale (HIV), per via oro-fecale (rotavirus) o mediante vettori animali (virus della rabbia).



Introduzione

SARS-CoV-2

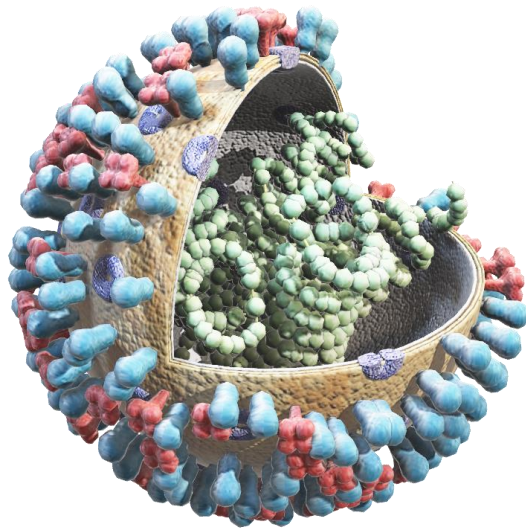
Il SARS-CoV-2 è un ceppo virale della specie coronavirus correlato alla SARS facente parte del genere Betacoronavirus. Settimo coronavirus riconosciuto in grado di infettare gli essere umani. I Betacoronavirus sono virus a RNA a filamento singolo positivo di cui i pipistrelli e i roditori sono considerati riserve virali.



Introduzione

SARS-CoV-2

Il SARS-CoV-2 è un ceppo virale della specie coronavirus correlato alla SARS facente parte del genere Betacoronavirus. Settimo coronavirus riconosciuto in grado di infettare gli esseri umani. I Betacoronavirus sono virus a RNA a filamento singolo positivo di cui i pipistrelli e i roditori sono considerati riserve virali.



Il quadro clinico è caratterizzato prevalentemente da **febbre**, tosse secca e **dispnea**. In misura minore si possono riscontrare sintomi gastrointestinali quali vomito e diarrea, manifestazioni oculari come la congiuntivite ed esantemi cutanei. Particolarmente distintiva risulta l'insorgenza di anosmia e disgeusia. Le principali complicanze associate includono **polmonite interstiziale severa**, sindrome da insufficienza multiorgano e, nei casi più critici, esito infausto.

La **trasmissione** avviene principalmente per via aerea attraverso goccioline respiratorie emesse durante la fonazione, la tosse o gli starnuti. Le strategie preventive fondamentali comprendono il mantenimento del distanziamento interpersonale, l'adozione di rigorose pratiche di igiene personale e ambientale, nonché l'utilizzo corretto dei dispositivi di protezione individuale, in particolare delle mascherine.

Introduzione



Mutazione dei Virus

Ogni processo evolutivo ha come base delle mutazioni che possono verificarsi casualmente ogni volta che una cellula o un virus si replica. Le mutazioni creano, all'interno di una popolazione, variazioni che consentono alla selezione naturale di amplificare i tratti che aiutano le creature a prosperare. I virus, anche se tecnicamente non considerati forme di vita, mutano ed evolvono infettando le cellule degli organismi ospite e replicandosi. Le modifiche del codice genetico del virus che ne derivano possono aiutarlo a passare più facilmente da un essere umano all'altro oppure ad eludere le difese del sistema immunitario. A questo proposito, è possibile distinguere:

VARIANTI DI PREOCCUPAZIONE


Presentano maggiore trasmissibilità, severità della malattia aumentata, riduzione dell'efficacia dei vaccini. Richiedono azioni immediate.

VARIANTI DI INTERESSE

Presentano mutazioni che potrebbero influenzare la trasmissibilità, la gravità della malattia o l'efficacia dei vaccini.

VARIANTI SOTTO MONITORAGGIO

Varianti con potenziale rischio, ma senza evidenze sufficienti per classificarle come VOI o VOC.



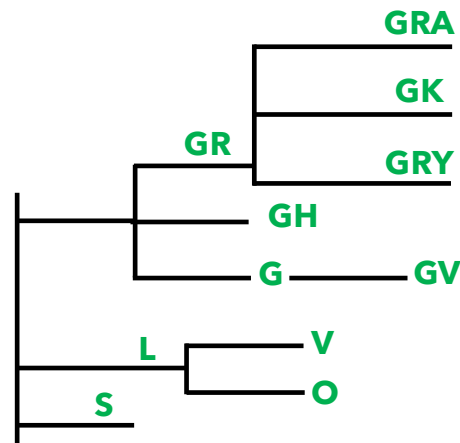
Introduzione

Varianti del Covid-19

A partire dalla fine del 2019, sono emerse numerose **varianti** di SARS-CoV-2, caratterizzate da mutazioni che, in alcuni casi, hanno determinato un incremento della trasmissibilità del virus o una modifica della risposta immunitaria dell'ospite.

La categorizzazione delle varianti avviene mediante sistemi di nomenclatura consolidati, tra cui **PANGO** e Nextstrain, che consentono il raggruppamento dei virus in cladi o lignaggi genetici sulla base delle mutazioni condivise.

Tra le varianti principali identificate figurano **Alfa**, **Beta**, **Gamma**, **Delta** e **Omicron**. Ciascuna di esse presenta specifiche mutazioni in grado di influenzare la trasmissibilità, il quadro sintomatologico e, in alcuni casi, la letalità dell'infezione.



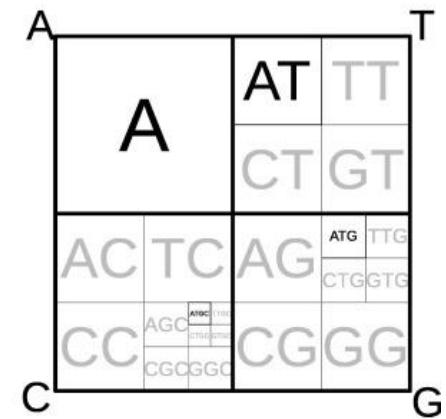
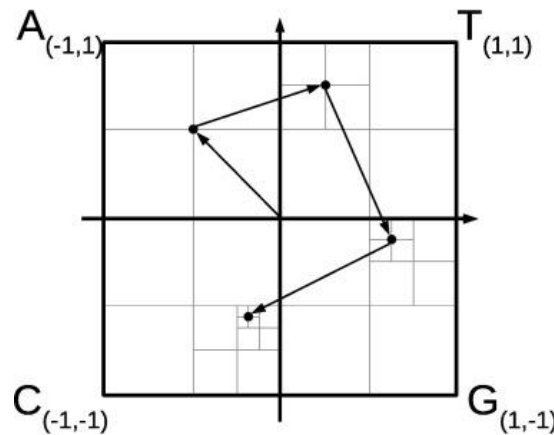
Introduzione

Algoritmi di codifica della sequenza amminoacidica

La **Chaos Game Representation** è un algoritmo computazionale utilizzato per rappresentare graficamente sequenze di acidi nucleici (DNA o RNA) in uno spazio bidimensionale. Ogni nucleotide (A, T/U, C, G) viene associato a un vertice di un quadrato, e la sequenza viene tracciata iterativamente calcolando il punto medio tra la posizione corrente e quella corrispondente al nucleotide successivo. Il risultato è una mappa visiva che riflette la composizione e la struttura della sequenza.

La presenza di **basi ambigue** o **mancanti** nelle sequenze può compromettere la precisione della rappresentazione. Per ovviare a tale problema, si ricorre all'utilizzo della matrice di **Frequenza CGR**.

Questo approccio permette di mantenere una rappresentazione informativa anche in presenza di ambiguità nella sequenza originaria.

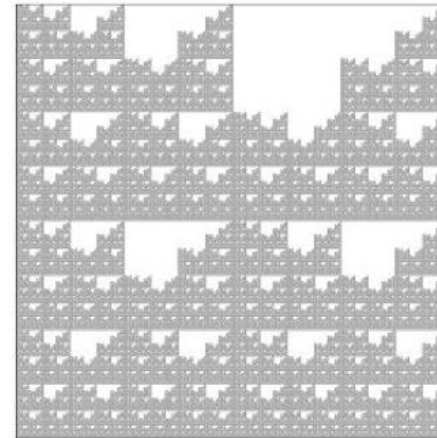
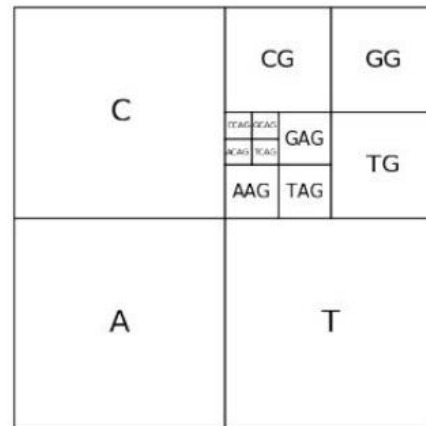


Introduzione

Algoritmi di codifica della sequenza amminoacidica

La **Frequency Chaos Game Representation** rappresenta un'estensione del metodo CGR, in cui si tiene conto della frequenza di occorrenza dei *k-mer* (sottosequenze di lunghezza *k*) all'interno di una sequenza genomica. Questo approccio consente la generazione di una matrice di densità che può essere visualizzata in forma bidimensionale o tridimensionale, offrendo una rappresentazione quantitativa della distribuzione dei motivi nucleotidici o amminoacidici.

L'FCGR si rivela particolarmente efficace nell'individuazione di pattern ricorrenti e specifici all'interno di sequenze di DNA, RNA o proteine, e trova applicazione in numerosi ambiti, tra cui la classificazione filogenetica, l'identificazione di specie, e l'analisi comparativa tra genomi.

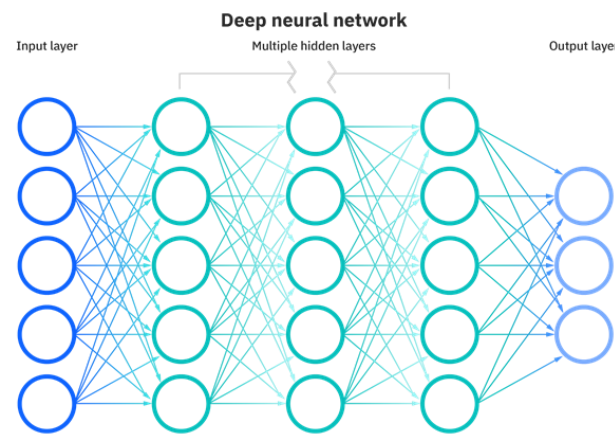


Introduzione

Deep Learning e Reti Neurali

Il **Deep Learning** rappresenta un sottoinsieme del Machine Learning che si basa sull'utilizzo di **reti neurali artificiali profonde**, costituite da numerosi strati interni, per l'analisi e l'apprendimento da grandi volumi di dati. Il principio di funzionamento prevede che un input, come ad esempio un'immagine, venga elaborato attraverso una serie di strati della rete, fino a produrre un output, come la classificazione dell'immagine stessa.

Le **architetture profonde** (Deep Neural Networks, DNN) permettono di modellare relazioni complesse tra i dati, migliorando sensibilmente le prestazioni in compiti quali la classificazione o il riconoscimento di pattern. Questo approccio ha trovato applicazione in numerosi ambiti, tra cui il **riconoscimento facciale**, la **classificazione automatica delle immagini**, e i **sistemi di guida autonoma**.

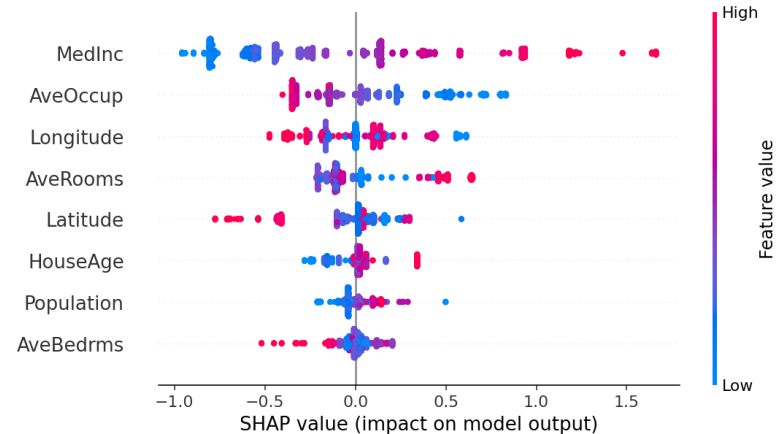


Introduzione

SHAP Values e Interpretabilità dei Modelli

Nel campo del **deep learning**, l'**interpretabilità dei modelli** rappresenta un aspetto fondamentale, in particolare nelle applicazioni biomediche, dove è essenziale comprendere quali caratteristiche dei dati influenzano le predizioni del modello. Questo requisito non è solo importante per motivi di trasparenza, ma anche per garantire l'affidabilità scientifica dei risultati ottenuti.

Una delle tecniche più diffuse per l'analisi interpretativa è rappresentata dai **valori SHAP** (*SHapley Additive exPlanations*), che si fondano sulla teoria dei giochi cooperativi. Questa metodologia assegna a ciascuna feature di input un valore che quantifica il suo contributo specifico alla predizione del modello, sia in termini positivi che negativi. In altre parole, i valori SHAP permettono di decomporre la decisione del modello in una somma di contributi attribuiti alle singole variabili, offrendo così una spiegazione chiara e localizzata delle previsioni.



Metodologia

Raccolta e processingamento dei Dati

Il processo di analisi si articola in diverse fasi, a partire dalla **raccolta e preprocessingamento dei dati**. Le sequenze genomiche utilizzate per questo studio provengono dalla piattaforma **GISAID**, un database di riferimento internazionale per la condivisione di sequenze del virus **SARS-CoV-2**. Sono stati adottati specifici criteri di selezione: sono state incluse solo sequenze **complete** (con lunghezza superiore a 29.000 basi), **ad alta copertura** (con una percentuale di nucleotidi ambigui compresa tra lo 0,05% e l'1%), e **provenienti da ospiti umani**. Inoltre, l'analisi si è focalizzata su due clade virali, **V** e **GK**, scelti per la loro rilevanza sia genetica che epidemiologica.

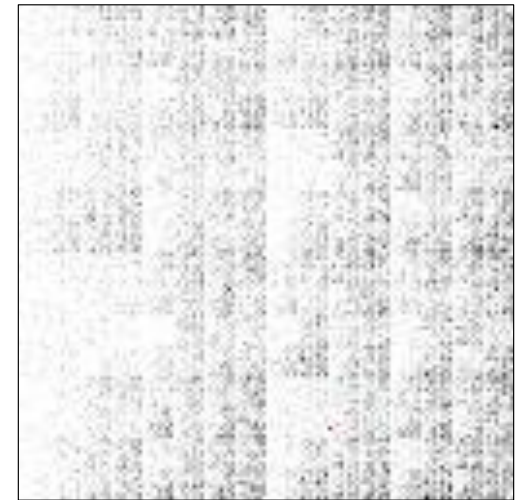
CLADE	NUMERO
V	4802
GK	5113

Metodologia

Generazione delle immagini

Successivamente, si è proceduto con la **generazione delle immagini** a partire dalle sequenze, utilizzando la tecnica della **Frequency Chaos Game Representation (FCGR)**.

Questo metodo consente una rappresentazione visiva avanzata del DNA, basata sulla frequenza di comparsa dei **k-mer**, ovvero sottosequenze di lunghezza k. I nucleotidi (A, C, G, T) vengono mappati all'interno di una griglia secondo uno schema quadripartito, dando origine a immagini in **scala di grigi**: le aree più scure indicano una maggiore frequenza di determinati k-mer, permettendo così una rappresentazione informativa delle caratteristiche genomiche.



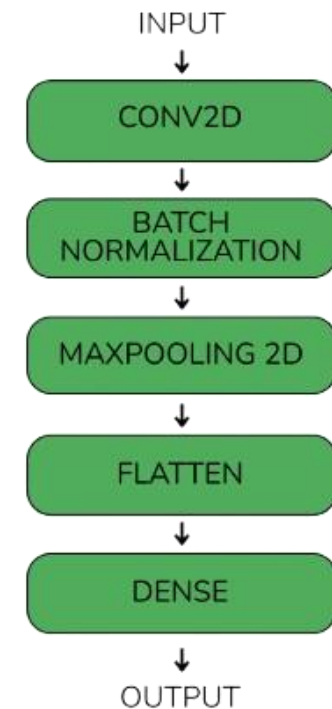
Metodologia

Progettazione e addestramento del modello

Il passo successivo ha riguardato la **progettazione e l'addestramento di un modello di deep learning**, in particolare una **rete neurale convoluzionale (CNN)**, utilizzata per classificare le immagini generate tramite FCGR. Le CNN sono particolarmente efficaci nell'identificare **pattern gerarchici** nei dati visivi grazie all'uso di **campi recettivi locali, pesi condivisi** e operazioni di **pooling**.

Il modello ha ricevuto in input immagini in scala di grigi di dimensione **128×128**, ed è stato strutturato con strati **Conv2D, MaxPooling2D, ReLU, e Batch Normalization**. L'output è una **classificazione binaria** tra il clade V e il clade GK, ottenuta tramite una funzione di attivazione **softmax**.

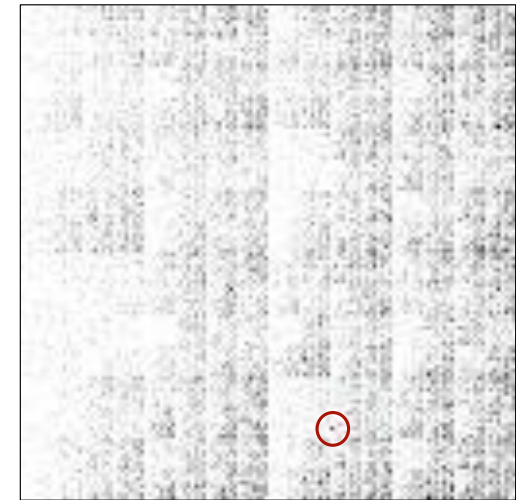
L'addestramento è stato eseguito utilizzando l'ottimizzatore **Adam**, la funzione di perdita **binary_crossentropy**, e un **batch size** pari a 64.



Metodologia

Interpretazione del modello tramite SHAP Values

Un elemento centrale dello studio è stata l'**interpretazione del modello** mediante l'impiego dei **valori SHAP** (*SHapley Additive exPlanations*), una tecnica ispirata alla teoria dei giochi che consente di attribuire a ciascun **pixel** dell'immagine (e quindi a ciascun k-mer) un valore che rappresenta il suo contributo alla predizione finale. L'**applicazione di SHAP** ha permesso di **visualizzare l'impatto di specifiche combinazioni nucleotidiche** sul risultato del modello: le aree evidenziate con **colori caldi** indicano le regioni più determinanti per distinguere tra i due clade. Questa capacità interpretativa è fondamentale per garantire **trasparenza** e **affidabilità** del modello, oltre che per **collegare le attivazioni neuronali a tratti genomici specifici**, offrendo un'opportunità concreta di analisi biologica.



Risultati

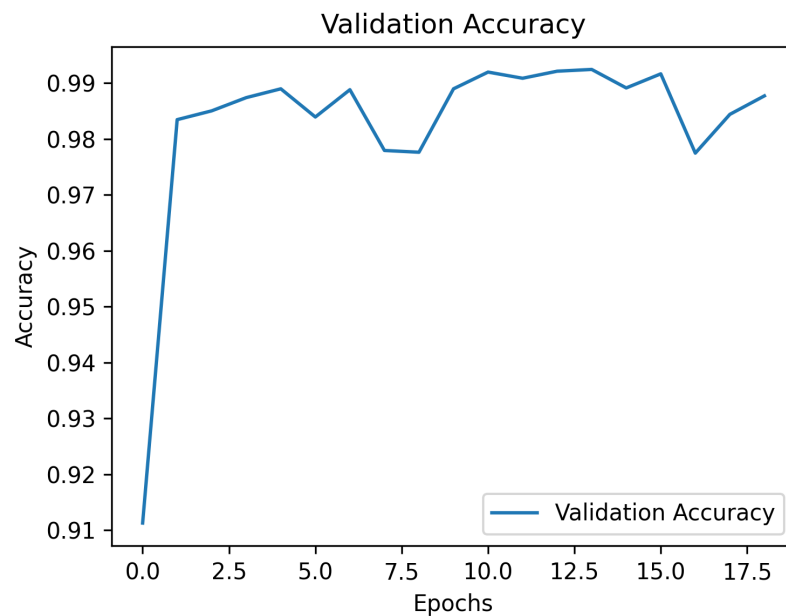
Obiettivi della sperimentazione e Configurazione

Lo scopo principale di questa sperimentazione è stato valutare l'efficacia di un modello di deep learning, basato su reti neurali convoluzionali (CNN), nella **classificazione delle sequenze genomiche del virus SARS-CoV-2**, con particolare riferimento alla distinzione tra i **clade V** e **GK**. Oltre alla performance predittiva, la sperimentazione ha indagato l'impatto delle scelte progettuali, quali il parametro k nella rappresentazione FCGR, l'architettura della rete e le tecniche di preprocessing, sull'efficacia complessiva del modello. Un ulteriore obiettivo ha riguardato l'interpretabilità, valutata tramite l'uso dei **valori SHAP**, al fine di comprendere le caratteristiche che guidano le decisioni del modello. Il parametro k , che determina la granularità della rappresentazione FCGR, è stato testato con due configurazioni: **$k=7$** e **$k=9$** . L'aumento di k permette una rappresentazione più dettagliata delle sequenze, influenzando la qualità visiva delle immagini e, di conseguenza, la precisione della classificazione.

Risultati

Analisi dei risultati

Con $k=7$, le immagini FCGR risultano di dimensione **128×128** , offrendo un buon equilibrio tra risoluzione e densità informativa. Il modello ha rapidamente raggiunto una validazione superiore al 98% in poche epoche, stabilizzandosi attorno al 99% senza segni evidenti di overfitting.



Test	Precision	Recall	F1-Score	Support
Clade V	0.90	0.99	0.95	1042
Clade GK	0.99	0.88	0.93	941

Risultati

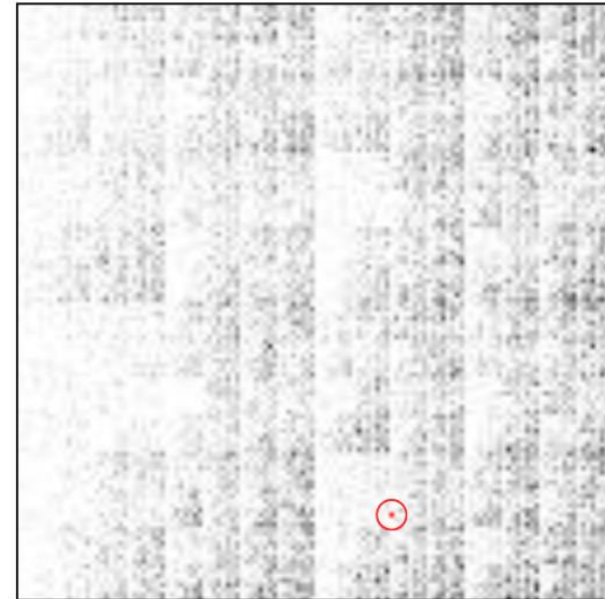
Analisi dei risultati

Con $k=7$, le immagini FCGR risultano di dimensione **128×128**, offrendo un buon equilibrio tra risoluzione e densità informativa. Il modello ha rapidamente raggiunto una validazione superiore al 98% in poche epoche, stabilizzandosi attorno al 99% senza segni evidenti di overfitting.

Clade V

La Sperimentazione effettuata, per $N=24, 50, 100$ e 200 ha prodotto 1 marker principale: **GGTTCAT**.

Applicando anche il Reverse & Complement non vi sono ottenuti miglioramenti.



Risultati

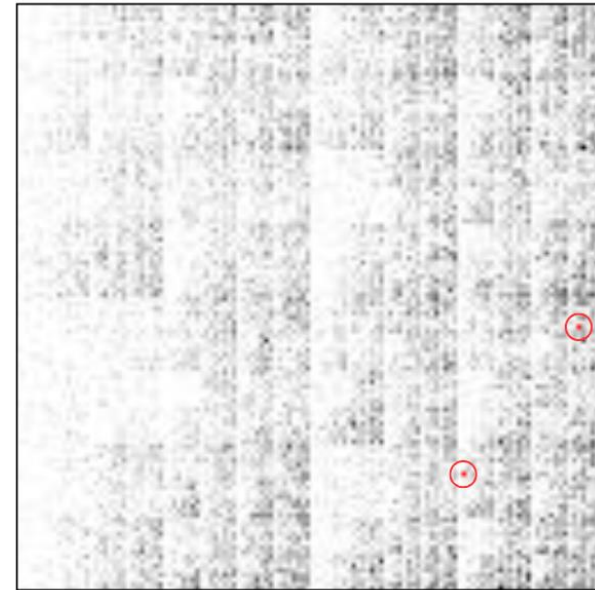
Analisi dei risultati

Con $k=7$, le immagini FCGR risultano di dimensione **128×128**, offrendo un buon equilibrio tra risoluzione e densità informativa. Il modello ha rapidamente raggiunto una validazione superiore al 98% in poche epoche, stabilizzandosi attorno al 99% senza segni evidenti di overfitting.

Clade GK

La Sperimentazione effettuata, per $N=24, 50, 100$ ha prodotto 1 marker principale: **ACACCTT**. Mentre, per $N=200$ evidenzia 1 marker in più: **TCAGGGT**.

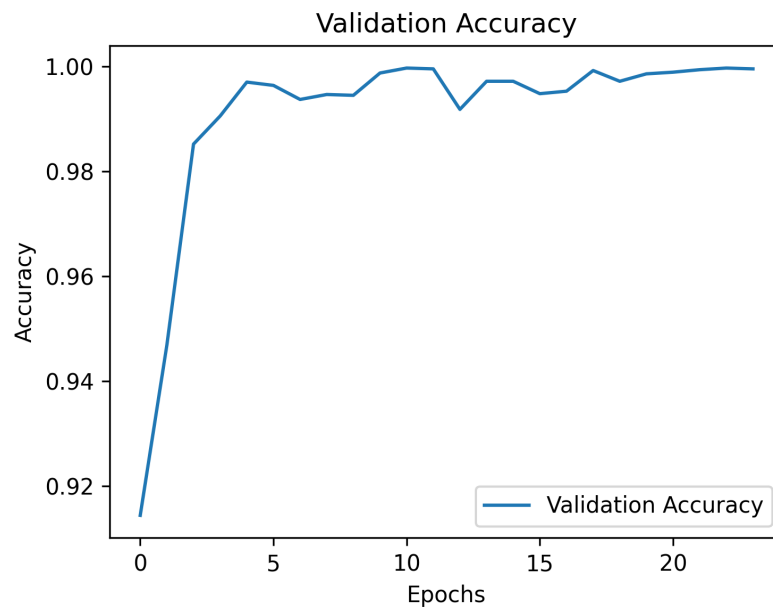
Applicando anche il Reverse & Complement per valori di $N=24, 50$ e 100 oltre a quello espresso vi è: **ACACCCT**. Mentre per $N=200$ si è aggiunto anche **TAACATC**.



Risultati

Analisi dei risultati

Con $k=9$, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.



Test	Precision	Recall	F1-Score	Support
Clade V	0.87	1	0.93	1042
Clade GK	1	0.84	0.91	941

Risultati

Analisi dei risultati

Con $k=9$, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.

Clade V

NON CI SONO RISULTATI

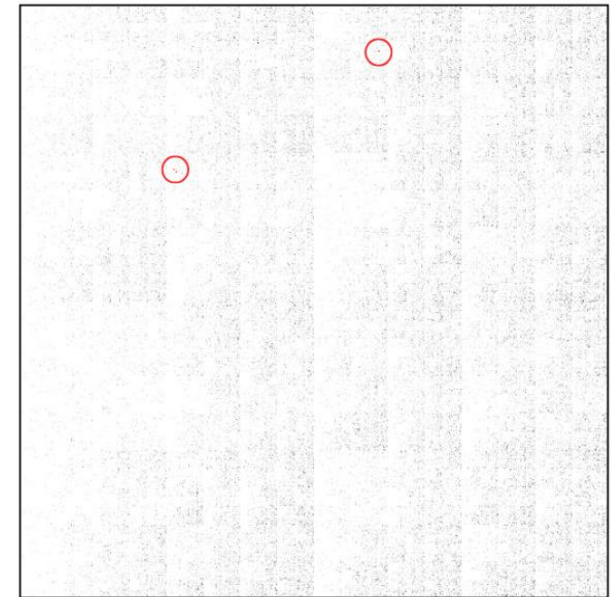
Risultati

Analisi dei risultati

Con $k=9$, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.

Clade GK

Per **$N=24$** emergono 3 marker principali:
TTTACCCTC, TCTACCCTC, TTTCGTCCG



Risultati

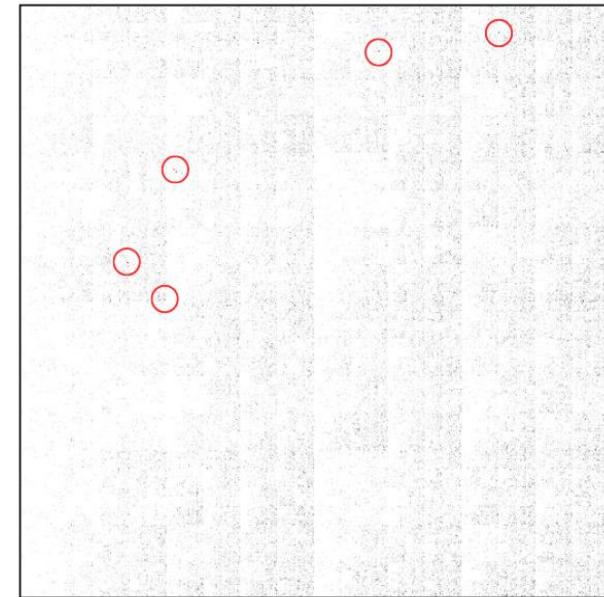
Analisi dei risultati

Con **k=9**, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.

Clade GK

Per **N=50** si osservano tre nuovi marker che si aggiungono ai precedenti:

TCTTTCTAC, TCTTTTTAC e TTTGTCCGG



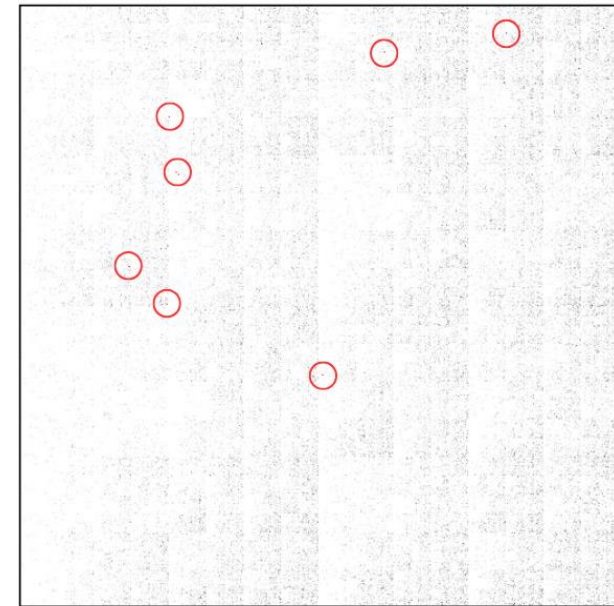
Risultati

Analisi dei risultati

Con **k=9**, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.

Clade GK

Per **N=100** vengono individuati due ulteriori marker: **AGCAAACCT** e **GTTTTGTCC**



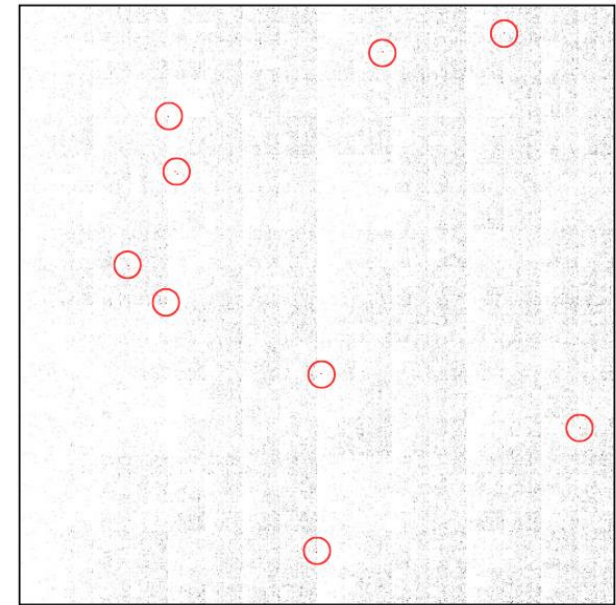
Risultati

Analisi dei risultati

Con **k=9**, le immagini risultano avere una dimensione di **512×512**, incrementando il dettaglio rappresentazionale. La curva di accuratezza ha mostrato una rapida crescita, con una validazione del 98% in tre epoche, stabilizzandosi prossima al 100% entro la quinta epoca.

Clade GK

Per **N=200** il modello riesce ad evidenziare altri due marker distintivi: **AGGGTGTTA** e **AAACCTTGT**, portando a dieci il numero complessivo di sequenze ricorrenti associate a questo clade.



Risultati

Confronti sui risultati

k=7: Il modello ha mostrato un buon equilibrio tra precisione e recall per entrambi i clade. In particolare, per il **Clade V**, si è osservata una precisione di 0.90 e una recall di 0.99, mentre per il **Clade GK**, la precisione è stata di 0.99 e la recall di 0.88.

k=9: Con l'aumento della lunghezza dei k-mer, si è notato un miglioramento della recall per il **Clade V**, che ha raggiunto un valore di 1.00, a fronte di una leggera diminuzione della precisione (0.87). Per il **Clade GK**, la precisione è rimasta elevata (1.00), ma la recall è diminuita a 0.84.



La lunghezza dei k-mer è un **compromesso** tra **precisione** e **recall**: valori più bassi offrono un equilibrio, mentre valori più alti possono migliorare una metrica a scapito dell'altra.



Non esiste una **lunghezza ottimale** universale per la classificazione delle sequenze, poiché le prestazioni dipendono dal metodo e dai dati.



La scelta della lunghezza dovrebbe quindi riflettere gli **obiettivi dell'analisi** e le caratteristiche dei dati, con k più basso per un equilibrio e k più alto per ottimizzare una metrica specifica.

Conclusioni

In conclusione, **la scelta del parametro k si conferma determinante** per il bilanciamento tra le metriche di valutazione. Non esiste un valore ottimale universale: la selezione va calibrata in base agli obiettivi dell'analisi e alla tipologia di dati. L'integrazione dei valori SHAP ha infine garantito un importante livello di interpretabilità, rendendo il modello non solo efficace, ma anche biologicamente significativo.

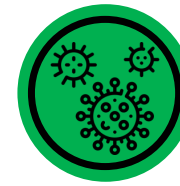
Sviluppi futuri



Per una migliore sperimentazione è consigliabile aumentare il numero di **K** fino a **19** o **21**.



Si potrebbero esplorare **architetture** di rete **più avanzate**, come modelli 3D, per catturare relazioni più complesse nelle sequenze genomiche.



Il metodo potrebbe essere **adattato ad altri ambiti**, come la classificazione di varianti patogene, l'analisi filogenetica o la predizione di resistenze farmacologiche

GRAZIE PER
L'ATTENZIONE

