UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

Corso di Laurea Magistrale in INFORMATICA

Corso Strumenti Formali per la Bioinformatica

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

Candidati:

Carmela Pia Senatore

Matricola 0522501721

Gennaro Capaldo

Matricola 0522507865

Anno Accademico 2023/2024

SOMMARIO

Nell'era dell'informazione genomica, l'analisi delle sequenze genomiche virali è diventata un aspetto cruciale per comprendere la diversità genetica del SARS-CoV-2 e per identificare le varianti emergenti di interesse epidemiologico. L'accurata classificazione delle varianti genomiche rappresenta una sfida fondamentale per la gestione e il contenimento della pandemia da COVID-19. In questo contesto, il nostro studio si propone di sviluppare un approccio innovativo per la classificazione delle sequenze genomiche del SARS-CoV-2 basato su tecniche avanzate di machine learning e rappresentazione del DNA tramite FCGR (Matrice di frequenza del chaos game).

L'obiettivo primario di questo lavoro è quello di ottenere una rappresentazione FCGR delle sequenze genomiche. Questa rappresentazione FCGR fornisce una base solida per l'applicazione di modelli di deep learning, in particolare le reti neurali convoluzionali (Convolutional Neural Networks, CNN), note per la loro capacità di catturare pattern complessi e caratteristiche all'interno delle immagini.

Successivamente, le sequenze genomiche così rappresentate vengono elaborate tramite una rete neurale convoluzionale complessa, progettata per la classificazione accurata delle clade. La CNN, addestrata su un ampio dataset di patients, è in grado di classificare le FCGR appartenenti alla proteina Spike o a altra classi con un tasso di accuratezza e F1-score elevati, ca. 98%.

INDICE

GLOSSARIO

DNA: Il DNA Sigla per acido desossiribonucleico, una grande molecola composta da nucleotidi a cui è affidata la codificazione delle informazioni genetiche; costituisce la sostanza fondamentale del gene ed è responsabile della trasmissione dei caratteri ereditari.

SEQUENZA GENOMICA: Successione ordinata delle basi azotate presenti nel genoma. Quattro basi azotate (adenina, timina, citosina, guanina) si susseguono, nel genoma di ogni organismo, secondo un ordine altamente preciso e ordinato e si raggruppano in diverse combinazioni a formare i geni, le unità di base dell'informazione genetica

FILE FASTA: Formato standard per rappresentare le sequenze biologiche. I file secondo questo formato hanno estensione .fa oppure .fasta e sono in plain text. Essi descrivono la sequenza primaria del genoma separata in righe di 60/80 bp e aggiungono delle informazioni addizionali come il cromosoma descritto, l'identificativo del genoma di riferimento e la sua localizzazione. Inoltre aggiunge un simbolo 1 se si tratta di una catena diretta di tipo 5'3' o-1 se di direzione 3'5'. Nato per il software FASTA che si occupa di allineamento locale. Sequenza genomica Sequenza di simboli sull'alfabeto a quattro lettere $\Sigma = \{A,T,G,C\}$ che rappresenta il sequenziamento di frammenti di DNA.

MAF: Un file MAF (Mutation Annotation Format) è un tipo di file utilizzato per descrivere mutazioni somatiche, principalmente in contesti di ricerca sul cancro. È spesso associato a progetti di genomica come il The Cancer Genome Atlas (TCGA), che mirano a mappare le mutazioni genetiche associate a diversi tipi di cancro. Il formato MAF è progettato per standardizzare la rappresentazione delle informazioni relative alle mutazioni, facilitando così la condivisione, l'analisi e l'interpretazione dei dati di sequenziamento del DNA.

VCF: Un file VCF (Variant Call Format) è un formato di file standardizzato utilizzato per memorizzare le informazioni relative alle varianti genetiche identificate rispetto a un genoma di riferimento. Questo formato è ampiamente adottato in bioinformatica per la condivisione e l'analisi di dati di varianti genetiche, come SNP (Single Nucleotide Polymorphisms), inserzioni, delezioni e altre varianti strutturali.

CGR : La rappresentazione del Caos dei Giochi Genetici (CGR, Chaos Game Representation) è una tecnica di visualizzazione che mappa sequenze di DNA, RNA o proteine in uno spazio geometrico in modo da rivelare pattern nascosti o proprietà strutturali delle sequenze stesse. Questo metodo è basato sul concetto matematico del gioco del caos, che utilizza iterazioni semplici per produrre figure complesse.

FCGR : La rappresentazione FCGR (Frequency Chaos Game Representation) è un metodo per visualizzare e analizzare sequenze di DNA utilizzando i principi del Chaos Game. Questo approccio trasforma le sequenze di nucleotidi in un'immagine che rappresenta la distribuzione e la frequenza delle varie basi (adenina, citosina, guanina e timina) all'interno della sequenza. L'FCGR mappa le sequenze di DNA in uno spazio quadridimensionale in modo che pattern e strutture all'interno delle sequenze diventino visivamente identificabili.

FRATTALI: I frattali sono strutture geometriche complesse che esibiscono un modello ripetitivo a diverse scale, noto come autosimilarità. In biologia e nel contesto bioinformatico, i frattali trovano applicazione nella descrizione e nell'analisi di strutture e processi che presentano pattern autosimilari o ripetitivi a livelli multipli di organizzazione. Questo concetto è particolarmente interessante

perché molti sistemi biologici mostrano proprietà frattali, dalla struttura dei polmoni e dei vasi sanguigni alla distribuzione dei cluster di cellule.

CROMOSOMA : Un cromosoma è una struttura organizzata di DNA e proteine che si trova nel nucleo delle cellule eucariotiche. I cromosomi svolgono un ruolo cruciale nella conservazione e nella trasmissione dell'informazione genetica da una generazione all'altra e sono fondamentali per i processi di divisione cellulare, inclusi la mitosi e la meiosi.

GENE: Un gene è una sequenza di DNA che contiene le istruzioni necessarie per costruire e mantenere le cellule di un organismo e per trasmettere le caratteristiche genetiche da una generazione all'altra. I geni sono unità funzionali dell'ereditarietà e codificano per proteine o RNA, che svolgono ruoli cruciali nei processi biologici. Ogni gene occupa una posizione specifica, o locus, su un cromosoma. La totalità dei geni di un organismo, insieme ad elementi regolatori non codificanti, costituisce il suo genoma. I geni influenzano una vasta gamma di tratti e condizioni, inclusi quelli fisici come il colore degli occhi e i gruppi sanguigni, nonché la suscettibilità a certe malattie.

MUTAZIONE: Una mutazione è un cambiamento nella sequenza del DNA di un organismo. Le mutazioni possono verificarsi in vari modi, come errori durante la replicazione del DNA, per esposizione a radiazioni o sostanze chimiche mutagene, o per infezioni virali. Possono variare in dimensione, dai cambiamenti in una singola base di DNA (mutazioni puntiformi) all'aggiunta o alla perdita di grandi segmenti di DNA o cromosomi interi.

CAPITOLO 1

INTRODUZIONE

1 Che cos’è? Informazioni sul covid-19

La pandemia di COVID-19, causata dal virus SARS-CoV-2, ha avuto un impatto senza precedenti sulla salute pubblica globale, influenzando profondamente la vita quotidiana, l'economia e i sistemi sanitari di tutto il mondo. 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. La capacità del SARS-CoV-2 di mutare e generare nuove varianti ha rappresentato una sfida continua per il controllo e la mitigazione della pandemia. Il genoma del SARS-CoV-2 è costituito da circa 30.000 basi di RNA, che codificano per diverse proteine strutturali e non strutturali. Tra queste, la proteina Spike (S) è di particolare interesse, in quanto è il principale mediatore dell'ingresso del virus nelle cellule umane e il bersaglio di gran parte dei vaccini sviluppati contro il COVID-19. Tuttavia, 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, una parziale resistenza alla risposta immunitaria, o una ridotta efficacia dei vaccini.

La sequenza genomica del SARS-CoV-2 è quindi un elemento cruciale per comprendere l'evoluzione del virus, monitorare la diffusione delle varianti e sviluppare strategie di controllo efficaci. Grazie ai progressi nella tecnologia di sequenziamento, oggi è possibile ottenere rapidamente e in modo relativamente economico le sequenze genomiche complete del virus da campioni clinici raccolti in tutto il mondo. Questo ha permesso la creazione di vaste banche dati globali contenenti milioni di sequenze genomiche del SARS-CoV-2, che forniscono una risorsa inestimabile per la sorveglianza e la ricerca.

L'analisi di queste sequenze genomiche è fondamentale per la classificazione delle varianti del SARS-CoV-2. Le varianti sono tipicamente identificate e classificate in base a mutazioni specifiche nella loro sequenza genetica, che possono alterare il comportamento del virus. Le varianti di interesse (VOI) e le varianti di preoccupazione (VOC) sono categorie stabilite dalle autorità sanitarie internazionali, come l'Organizzazione Mondiale della Sanità (OMS), per monitorare le mutazioni che potrebbero avere un impatto significativo sulla salute pubblica. Queste varianti possono influenzare la trasmissibilità del virus, la gravità della malattia, l'efficacia dei vaccini e la risposta ai trattamenti.

La classificazione accurata delle varianti è quindi essenziale per guidare le decisioni cliniche, le politiche sanitarie e le strategie di vaccinazione. Inoltre, la continua sorveglianza genomica permette di rilevare tempestivamente l'emergere di nuove varianti, facilitando interventi rapidi per prevenire la loro diffusione. In questo contesto, l'uso di tecnologie avanzate come il machine learning e le reti neurali convoluzionali sta diventando sempre più importante per analizzare e interpretare i vasti dati genomici disponibili, consentendo una classificazione più efficiente e precisa delle varianti virali.

Il progetto mira a esplorare la potenzialità della sinergia tra FCGR e CNN per una veloce e corretta classificazione, oltre che identificazione dei k-mers più importanti.

1.1 Classificazione delle varianti Covid-19

La classificazione delle varianti del SARS-CoV-2, il virus responsabile del COVID-19, è diventata una componente centrale nella gestione della pandemia. Con l'evoluzione continua del virus e l'emergere di nuove mutazioni, la necessità di monitorare e classificare queste varianti è essenziale per la salute pubblica, la ricerca scientifica e la risposta globale alla crisi. Le varianti del SARS-CoV-2 emergono a causa di mutazioni nel genoma del virus. Queste mutazioni possono essere singoli cambiamenti nucleotidici (sostituzioni), inserzioni, delezioni o ricombinazioni che alterano la sequenza genetica originale. Non tutte le mutazioni hanno un impatto significativo, ma alcune possono influenzare la trasmissibilità del virus, la gravità della malattia, la capacità del virus di sfuggire alla risposta immunitaria, o l'efficacia dei vaccini e dei trattamenti terapeutici.

Le autorità sanitarie internazionali, come l'Organizzazione Mondiale della Sanità (OMS) e i Centri per il Controllo e la Prevenzione delle Malattie (CDC), hanno sviluppato un sistema di classificazione delle varianti basato sul loro potenziale impatto sulla salute pubblica. Le varianti sono categorizzate in:

- **Varianti di Interesse (VOI, Variants of Interest)**: 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. Le VOI sono sottoposte a monitoraggio in quanto potrebbero evolvere in varianti più preoccupanti.
- **Varianti di Preoccupazione (VOC, Variants of Concern)**: Queste varianti presentano evidenze di una maggiore trasmissibilità, una severità della malattia aumentata, una significativa riduzione dell'efficacia dei vaccini, o una ridotta capacità di neutralizzazione da parte di anticorpi derivanti da infezioni precedenti o dalla vaccinazione. Le VOC richiedono azioni immediate a livello di salute pubblica, come l'intensificazione della sorveglianza, la modifica dei vaccini o l'introduzione di nuove misure di controllo.
- **Varianti sotto Monitoraggio (VUM, Variants under Monitoring)**: Sono 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. Queste varianti vengono attentamente monitorate per eventuali segnali che possano indicare un impatto significativo.

Il sequenziamento genomico è lo strumento principale utilizzato per identificare e classificare le varianti del SARS-CoV-2. 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, dove vengono analizzate e tracciate per monitorare la diffusione globale delle varianti.

La classificazione delle varianti del SARS-CoV-2 ha implicazioni dirette sulle strategie di salute pubblica e sulla gestione della pandemia. Identificare una variante come VOC può portare a restrizioni di viaggio, potenziamento della sorveglianza genomica, e modifiche nei protocolli di trattamento e vaccinazione. Per esempio, la scoperta di varianti come la Delta e la Omicron ha portato a

cambiamenti significativi nelle raccomandazioni per i vaccini, nelle strategie di testing e nelle misure di prevenzione a livello globale.

Inoltre, la classificazione delle varianti guida la ricerca scientifica su nuovi vaccini e trattamenti. Le mutazioni che influenzano la proteina Spike, ad esempio, possono richiedere lo sviluppo di vaccini aggiornati o adattati per mantenere l'efficacia contro queste nuove varianti.

Oltre al sequenziamento tradizionale, le tecniche di machine learning e l'intelligenza artificiale stanno giocando un ruolo crescente nella classificazione delle varianti. Modelli avanzati, come le reti neurali convoluzionali, vengono utilizzati per analizzare grandi quantità di dati genomici e identificare pattern che potrebbero non essere immediatamente evidenti con i metodi tradizionali. Questi strumenti permettono una classificazione più rapida ed efficiente delle varianti, facilitando l'identificazione precoce di varianti emergenti che potrebbero rappresentare nuove minacce.

2 Problema

Nel contesto della continua evoluzione del SARS-CoV-2, la classificazione delle varianti virali è diventata un elemento cruciale per la comprensione della pandemia di COVID-19 e per lo sviluppo di strategie di controllo efficaci. In particolare, il nostro progetto si concentrerà sull'analisi approfondita delle varianti appartenenti alle classi O e S, due categorie che rappresentano aspetti specifici dell'evoluzione genetica del virus.

Classe O (Originale o "Altro")

La **classe O**, spesso indicata come "Originale" o "Altro," raggruppa le varianti del SARS-CoV-2 che non rientrano nelle categorie di varianti di interesse (VOI) o varianti di preoccupazione (VOC) identificate dalle autorità sanitarie. Questa classe include le varianti che sono più vicine alla sequenza ancestrale del virus, ossia quelle sequenze che sono rimaste relativamente stabili o che presentano mutazioni che non hanno avuto un impatto significativo sulla trasmissibilità, sulla patogenicità o sull'evasione immunitaria.

L'analisi delle varianti della classe O è fondamentale per comprendere la base genetica del SARS-CoV-2 e per tracciare l'evoluzione del virus nel tempo. Queste varianti fungono da riferimento per lo studio delle mutazioni emergenti e delle loro implicazioni. Nonostante non rappresentino un rischio immediato o diretto come le VOC, le varianti della classe O offrono una finestra sull'evoluzione molecolare del virus e sulle possibili direzioni future del suo sviluppo.

Classe S (Spike)

La **classe S** si concentra sulle varianti che presentano mutazioni significative nella proteina Spike (S) del SARS-CoV-2, 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.

Le mutazioni nella proteina Spike possono alterare in modo significativo il comportamento del virus, influenzando la sua capacità di infettare le cellule, la sua trasmissibilità, e la sua capacità di sfuggire alla risposta immunitaria. Per esempio, mutazioni come la D614G, E484K, e N501Y, che sono state osservate in varianti preoccupanti come Alfa, Beta, Gamma e Delta, sono state associate a cambiamenti nell'affinità di legame al recettore ACE2 e alla neutralizzazione da parte degli anticorpi.

Il nostro progetto si propone di 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. Per ogni paziente si otterrà una rappresentazione FCGR, con k=7. Utilizzeremo tecniche avanzate di machine learning, come le reti neurali convoluzionali, per identificare e classificare queste varianti, con un particolare focus sui k-mer più importanti (k=7).

3 Struttura dello studio

Il presente studio è organizzato in modo da fornire una panoramica completa e dettagliata delle metodologie utilizzate, dell'applicativo sviluppato, dell'analisi dei risultati ottenuti e delle conclusioni derivanti dal lavoro svolto. La struttura del documento è suddivisa nei seguenti capitoli:

Capitolo 2: Metodologia Utilizzata

In questo capitolo, vengono descritte in dettaglio le metodologie utilizzate per l'analisi delle sequenze genomiche del SARS-Covid e dei metadati relativi ai pazienti. Viene introdotto il concetto di Teoria del chaos per la rappresentazione del DNA e la sua relativa matrice (FCGR), spiegando come questa rappresentazione permette di codificare le sequenze nucleotidiche in una forma che ne preserva le caratteristiche strutturali e funzionali rilevanti. Si discutono anche le tecniche di machine learning applicate, con particolare attenzione alle reti neurali convoluzionali (Convolutional Neural Networks, CNN), che sono state impiegate per la classificazione delle varianti.

Capitolo 3: Presentazione dei dati e discussione dei risultati

In questo capitolo, vengono presentati i dati utilizzati e la struttura dell'applicativo sviluppato per l'analisi delle sequenze genomiche del SARS-CoV-2. La prima sottosezione si occupa di fornire la descrizione del dataset, fase di data cleaning e analisi statistiche utilizzate per descrivere la coorte.

Il secondo paragrafo , invece, si occupa si svolgere l' analisi esplorativa dei dati , con particolare attenzione alle proteine e mutazioni.

Nella terza sottosezione, viene fornita una descrizione dettagliata dell'architettura software, illustrando i moduli principali e le loro interazioni. Si analizzano le scelte progettuali, inclusi gli algoritmi implementati e le librerie software utilizzate per il machine learning e l'elaborazione dei dati genomici.

Particolare enfasi viene posta sull'analisi della rete neurale convoluzionale impiegata. Si discutono l'architettura della rete, il numero di livelli, le funzioni di attivazione, e le tecniche di regolarizzazione utilizzate per prevenire l'overfitting. Inoltre, viene spiegato come la rete è stata addestrata per identificare pattern significativi nelle sequenze genomiche.

Nella quarta sezione, infine, si supportano i risultati principali di questo studio con particolare attenzione agli esperimenti svolti.

Capitolo 4: Conclusioni e Prospettive Future

Nel capitolo finale, si traggono le conclusioni dello studio, sintetizzando le principali scoperte e il loro impatto sulla comprensione delle varianti del SARS-CoV-2. Viene valutata l'efficacia del metodo FCGR e della rete neurale convoluzionale nella classificazione delle varianti, con particolare attenzione alle classi O e S.

Inoltre, vengono discusse le implicazioni di questi risultati per la sorveglianza genomica del SARS-CoV-2 e per lo sviluppo di future strategie di intervento. Si propongono possibili miglioramenti e direzioni future per la ricerca, tra cui

l'integrazione di ulteriori dati genomici e l'uso di tecniche avanzate di interpretazione del machine learning.

Infine, il capitolo conclude con una bibliografia completa che include tutte le fonti e i riferimenti scientifici utilizzati nel corso dello studio, garantendo una solida base documentale per i risultati presentati.

CAPITOLO 2

METODOLOGIA UTILIZZATA

Vengono descritte le basi teoriche che riguardano l'ambito biologico, statistico e informatico necessarie per l'analisi esplorativa di dati, la costruzione delle rappresentazioni FCGR e l'applicazione dei modelli di machine learning con successiva interpretazione.

1 Analisi esplorativa dei dati

L'analisi esplorativa dei dati comprende non solo la valutazione del grado di associazione tra variabili ma anche l'implementazione di curve di sopravvivenza al fine di avere una panoramica generale dei dati. In generale, l'obiettivo di questo tipo di analisi è quello di comprendere come le variabili influenzano o sono influenzate dalle altre, cogliere al meglio i fenomeni naturali e di identificare eventuali modelli o relazioni che possono esistere fra di esse.

Esistono diversi metodi per studiare le associazioni tra le variabili, per questo studio ci si avvale:

- La *statistica descrittiva*: un insieme di tecniche per descrivere e analizzare i dati in modo sintetico. Attraverso la statistica descrittiva, è possibile ottenere rappresentazioni grafiche quali istogrammi, box plot, e grafici a dispersione, che facilitano la comprensione della distribuzione e delle relazioni tra le variabili.
- I *modelli di sopravvivenza* si concentrano sull'analisi dei tempi fino all'occorrenza di un evento di interesse. Attraverso l'utilizzo di curve di sopravvivenza, è possibile studiare l'effetto di variabili esplicative sul tempo di sopravvivenza.

L'uso combinato di questi metodi permette di ottenere una comprensione approfondita delle relazioni tra le variabili e di selezionare in modo informato le caratteristiche più rilevanti.

1.1 Test del Chi-Quadro

Il *test del Chi quadrato* (χ^2) è un test statistico utilizzato per determinare se esiste una significativa associazione tra due variabili categoriche. È ampiamente utilizzato in statistica per testare ipotesi riguardanti la distribuzione di frequenze osservate in diverse categorie, confrontandole con le frequenze teoriche che ci si aspetterebbe se le variabili fossero indipendenti.

Si organizzano i dati in una tabella di contingenza, con una dimensione rappresentante le categorie di una variabile e l'altra le categorie della seconda variabile. Per *tabella di contingenza* si intende una tabella che permette di analizzare la relazione tra due o più variabili, che possono essere di tipo quantitativo discreto o qualitativo. [1] Indichiamo con n_{ij} le frequenze assolute relative alle variabili di riferimento. La tabella di contingenza sarà quindi rappresentata da *r* righe e *s* colonne.

Le frequenze assolute così individuate soddisferanno le seguenti relazioni:

$$n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$$

$$n_{i.} = \sum_{j=1}^{s} n_{ij}$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

dove si ha rispettivamente la somma delle frequenze assolute:
-totale è *n*;
-della riga *i*-esima è $n_{i.}$;
-della colonna *j*-esima è $n_{.j}$.
Di conseguenza sarà possibile indicare le frequenze relative e marginali come segue:

$$f_{ij} = \frac{n_{ij}}{n}; \qquad f_{i.} = \frac{n_{i.}}{n}; \qquad f_{.j} = \frac{n_{.j}}{n}.$$

Le frequenze marginali delle corrispondenti colonne corrispondono alla *j*-esima componente del *centro di gravità della nuvola* dei profili riga, analogamente può dirsi dei profili colonna: in entrambi i casi il centro di gravità è dato dal profilo marginale.
L'analisi della tabella di contingenza, analizzata tramite distribuzioni di frequenze relative, consente di rendere confrontabili le diverse modalità di una stessa variabile.
Si definiscono le ipotesi del test, per cui l'ipotesi nulla (H_0) asserisce che non esiste una relazione significativa tra le due variabili, sono indipendenti. L'ipotesi alternativa (H_1) ipotizza che esiste una relazione significativa tra le due variabili; non sono indipendenti.
Si assumono le frequenze attese per ciascuna combinazione di categorie delle due variabili, basandosi sull'ipotesi che le variabili siano indipendenti. La frequenza attesa per ogni cella si calcola come:

$$E_{ij} = \frac{(Totale\ riga_i)\ x\ (Totale\ colonna_j)}{Totale\ generale}$$

Il valore del Chi quadrato si calcola come la somma delle differenze al quadrato tra le frequenze osservate e quelle attese, divise per le frequenze attese, per tutte le celle della tabella. [2]

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Dove O_i sono le frequenze osservate e E_i quelle attese. I gradi di libertà (df) per il test del Chi quadrato sono determinati da:

$$df=(i-1)\times(j-1)$$

Con *i* il numero di righe e *j* il numero di colonne della tabella di contingenza.

Il valore calcolato del Chi quadrato si confronta con un valore critico dalla distribuzione del Chi quadrato, che dipende dal numero di gradi di libertà e dal livello di significatività desiderato (ad esempio, 0.05). Se il valore calcolato supera il valore critico, si rifiuta l'ipotesi nulla di indipendenza tra le variabili, per cui le due variabili sono associate. Il rifiuto dell'ipotesi nulla implica che esiste una relazione statistica significativa tra le due variabili, suggerendo che non sono indipendenti. E' , tuttavia, necessario tenere conto che il test del Chi quadrato non fornisce informazioni sulla direzione o sulla natura della relazione, solo sulla sua esistenza.

2 Chaos Game Representation (CGR)

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 unico e la sequenza fonte può essere recuperata dalle coordinate in modo tale che la distanza tre le posizioni misura la somiglianza tra le sequenze corrispondenti. Trovare schemi nascosti in lunghe sequenze può essere difficile e laborioso. Rappresentare queste sequenze in modo visivo può spesso aiutare. [7] L'algoritmo Chaos Game è stato sviluppato da Barnsley [4] per costruire frattali basati su input casuali e fu successivamente esteso al DNA come input da Jeffrey [5]. I risultati vengono richiamati Chaos Game Representation (CGR). Con l'avvento del frattale geometrico di Mandelbrot negli anni '70 , emersero degli algoritmi multipli per costruire frattali. I frattali sono ricorsivi, in scala modelli invarianti e la loro dimensione è una frazione . Esistono diversi algoritmi per costruire frattali, come i Sistemi L, e Poteri di Kronecker.

L'algoritmo di Barnsley iterazione casuale per creare immagini di frattali utilizza l'insieme di attrazione fisso di un sistema di funzioni iterate (IFS). Prima di definire IFS, è necessario fissare alcuni concetti preliminari. [3]

Trasformazione affine nello spazio euclideo. E' possibile scrivere una trasformazione *affine* bidimensionale nel piano Euclideo $w: R \to R^2$ nella forma:

(1) $$w(x_1, x_2) = (ax_1 + bx_2 + e, cx_1 + dx_2 + f)$$

Dove *a,b,c,d,e,f,* sono numeri reali. Si utilizza quindi la notazione compatta:

(2) $$w(x_1, x_2) = w\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$

$$= Ax + t$$

La matrice A è una matrice reale A può essere riscritta nella forma:

(3) $$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} r_1 \cos\theta_1 & -r_2 \sin\theta_2 \\ r_1 \sin\theta_1 & r_2 \cos\theta_2 \end{bmatrix}$$

Un esempio di *trasformazione lineare è :*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (4)$$

Se si vuole rapportare questo risultato in termini geometrici, la forma della figura rimane la stessa ma cambia la taglia e di rotazione.

Iterated Function System. Un sistema di funzioni iterate è un insieme di mappature contratte in uno spazio metrico. Questo può essere utilizzato per creare immagini dei *frattali.* Viene illustrato l'algoritmo tramite un esempio: Si considerino dei maps espressi come: w_1, w_2, w_3 , ciascuno con probabilità di fattore (1/3). Questi valori possono essere anche valutati attraverso una tabella dove l'ordine dei coefficienti da *a* a *f* corrisponde all'ordine presentato *nell'equazione*. Per creare l'immagine frattale usando IFS, si sceglie un punto di partenza (x_0, y_0). Successivamente, si sceglie in maniera randomica una mappa per l'IFS e si inizia a valutare dal punto di partenza (x_0, y_0) sino al prossimo punto (x_1, y_1). Lo si ripete molte volte finché il pattern non diventa visibile. La rappresentazione grafica è proprio il Triangolo Sierpinski con i vertici locati in (0,0), (0,1) e (1,1).

L'algoritmo alla base della CGR consiste nel mappare una sequenza, ovvero una rappresentazione 1D su uno spazio dimensionale superiore, tipicamente allo spazio 2D [4]. Originariamente era stato sviluppato per costruire il triangolo di Sierpinski. A tal fine, ai vertici di un triangolo vengono assegnati i numeri da uno a tre. Basandosi su un punto iniziale selezionato casualmente (S), un vertice viene scelto casualmente (V1) e viene disegnato un punto P1 a metà della distanza dal vertice V1. Questo processo viene ripetuto, con P1 come nuovo punto di partenza. Il secondo punto (P2) viene disegnato a metà verso il secondo vertice selezionato casualmente (V2). Ripetendo questo algoritmo, emerge il triangolo di Sierpinski (vedi *Fig. 1* a destra). Il CGR viene presentato come un grafico a dispersione (più frequentemente quadrato), in cui ogni angolo rappresenta un elemento che appare nella sequenza. I risultati di questi CGR possono sembrare molto simili a frattali, ma anche così possono essere visivamente riconoscibili e distinguibili. Le caratteristiche distintive possono essere rese quantitative, con un concetto di distanza tra le immagini. [3]

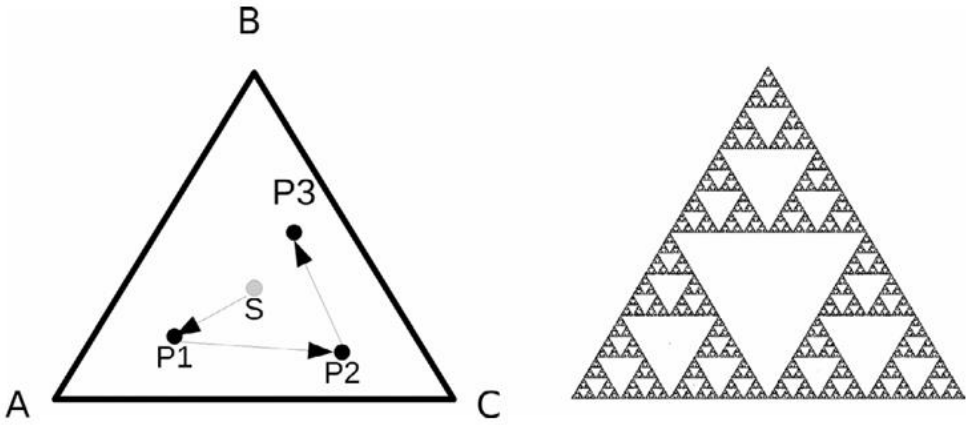


Figura 1: Chaos Game Algorithm

Barnsley [4] descrive il CGR come un sistema di funzioni iterate (IFS), che è basato sulla teoria degli insiemi. Negli ultimi anni, una nota più compatta è stata stabilita una soluzione basata su un approccio geometrico :

$$P_i^j = P_{i-1}^j + sf(V_{i-1}^j - P_{i-1}^j)$$

Nel 1990, Jeffrey [5] propose l'applicazione dell'algoritmo CGR al DNA, portando ad ampie applicazioni in bioinformatica. La novità stava nell'utilizzare questa rappresentazione come un modo nuovo per visualizzare le sequenze di nucleotidi. Tutti gli altri hanno esteso il CGR alle proteine, a causa della sua proprietà di unicità:

- una sequenza è rappresentata come un modello unico
- una sequenza viene mappata su coordinate univoche
- un CGR mappa tutte le possibili sequenze in tutte le possibili lunghezze in 2D o in uno spazio 3D
- una singola coordinata codifica l'input della sequenza completa
- il punto di partenza influenza di poco il risultato

Jeffrey [5] è stato il primo ad applicare la CGR al DNA. 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 [5] ha osservato che per sequenze casuali non emergono modelli visibili (vedi Fig. 3 in alto a sinistra). CGR è stato anche utilizzato per esaminare visualmente la qualità dei generatori di numeri casuali. [6] Per generatori di numeri casuali di alta qualità non emergono schemi visibili o pattern nella rappresentazione CGR. Questa caratteristica di CGR ha portato a due applicazioni, vale a dire strumenti per l'analisi visiva dei dati, ad esempio, visualizzare e analizzare generatori di numeri pseudocasuali e per confronti di sequenze senza allineamento. Si utilizza un CGR di forma quadrata [3], in cui i quattro angoli hanno il nome di ciascuna base. In Fig. 3, 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). Originariamente Jeffrey utilizzava a notazione diversa con un CGR compreso tra (0,0) e (1,1), con una diversa assegnazione dei vertici. Ci sono diversi modi per assegnare i nucleotidi ai vertici, cioè tre diversi orientamenti più rotazioni e specchiature. Secondo Jeffrey, ogni punto della rappresentazione CGR corrisponde esattamente a una sottosequenza (a partire dalla prima base), fino alla risoluzione dello schermo. Pertanto questo motivo grafico indica modelli ripetuti nella sequenza genetica. 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.

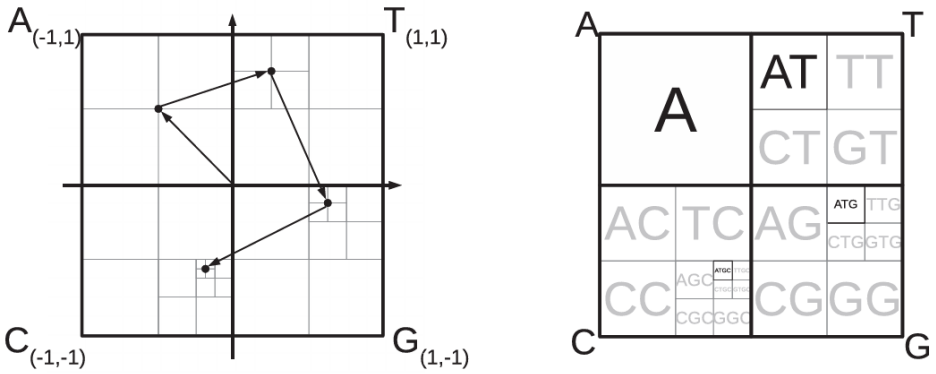


Fig. 2. Chaos Game Representation and algorithm for DNA. Left: CGR algorithm for four vertices with corresponding labels, i.e., A, C, G, and T, as well as corresponding coordinates. In CGR the center has the coordinates (0,0) and the CGR spans from (−1,−1) to (1,1). Right: Division of the CGR space due to the iterative process.

Figura 2: Chaos Game rappresentazione e algoritmo per il DNA.

L'immagine completa presente nella figura sottostante ha un pattern *frattale identificabile.* 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. Per comprendere a pieno il *double scoop pattern* è necessario comprendere il vero e proprio significato del CGR. Ogni punto tracciato nel CGR corrisponde a una base, e a seconda di dove è posizionato, si può risalire e risalire a parti della sequenza che si sta esaminando. La figura 3.a mostra l'area corrispondente in relazione del CGR e la sequenza del DNA. In riferimento a questa figura, si può vedere che per ogni punto che corrisponde alla base G, si troverà nel quadrante in alto a destra della trama CGR. Per vedere qual è la base precedente, si può dividere il quadrante in sottoquadranti (etichettandoli nello stesso ordine dei quadranti) e a seconda di dove si trova il punto, si può determinare qual è la base precedente della sequenza. Si può ripetere questo step all'infinito per trovare l'ordine delle basi che appaiono nelle sequenze. La Figura 3.b mostra il quadrato CGR dove tutti i quadranti sono non pieni.

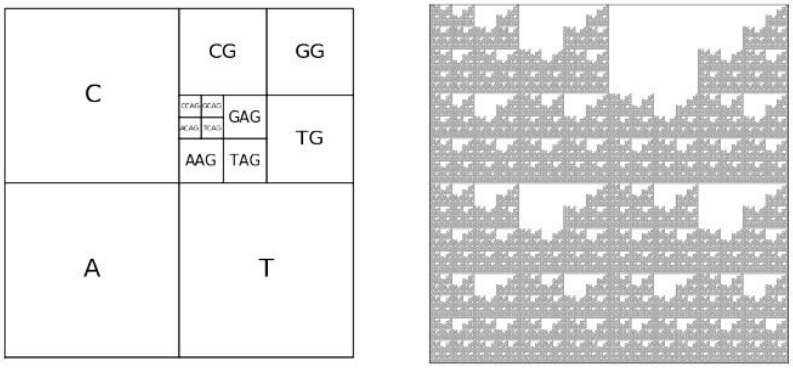


Figura 3: a) corrispondenza tra sequenze DNA e area del CGR della sequenza DNA; b) spiegazione del doppio scoop pattern plot del quadrato CGR.

Proprietà di una rappresentazione CGR di una sequenza DNA. Si possono però notare alcune proprietà fondamentali.

a. 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.

b. Pertanto qualsiasi modello visibile nel CGR corrisponde a qualche schema nella sequenza delle basi.

c. La risoluzione dello schermo limita i dettagli e vale per qualsiasi CGR. Tuttavia, come con tutti i frattali, compresi quelli generati dai codici IFS, qualsiasi parte dell'immagine può essere ingrandita, rivelando una struttura più fine. Quindi, se è presente un'area di interesse in cui la struttura sospetta è oscurata, può essere ingrandita per mostrare la struttura fine dei punti e, quindi, la struttura delle sequenze che producono i punti. Questo ingrandimento è illimitato (purché ci siano più basi nella sequenza)

d. Basi adiacenti nella sequenza non vengono plottate in maniera adiacente l'un l'altra (a eccezione di quando il primo punto è vicino all'angolo e lo stesso vale per la stessa base); 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.

e. La questione che quando due punti sono vicini nella rappresentazione CGR rappresentano sequenza simili è più complesso. In generale, due punti vicini potrebbero corrispondere a differenti sequenze. Tuttavia, questa situazione può verificarsi solo se i due punti, pur essendo vicini, si trovano in quadranti diversi della foto. Poiché una base viene sempre tracciata nel suo quadrante, qualsiasi sequenza sarà sempre tracciata da qualche parte nel quadrante della sua ultima base, e viceversa due punti qualsiasi nello stesso quadrante devono avere la stessa ultima base. Inoltre, la nozione di quadrante è ricorsiva; ogni quadrante può essere diviso in quadranti, ecc. Inoltre, sempre a causa del fatto che nel suo quadrante è tracciata una base, vale anche il contrario: se due punti sono all'interno dello stesso quadrante, corrispondono a successioni con la stessa ultima base; se si trovano nello stesso sotto quadrante le sequenze hanno le stesse ultime due basi; se sono nello stesso sub-sub-quadrante hanno le stesse ultime tre basi, ecc. E' possibile quindi, sulla base delle ultime affermazioni, definire il teorema 1:

Teorema 1. In una rappresentazione CGR il cui lato è di lunghezza 1, due sequenze con suffisso di lunghezza k sono contenuti all'interno del quadrato di lato di lunghezza 2^{-k}. Inoltre, il centro del quadrato è definito sulla base delle seguenti definizioni ricorsive:

(a) Il centro del suffisso 0 è (1/2, ½)

(b) Il centro del quadrato contenente sequenze con suffisso in *w* è in *(x,y),* per cui:

i. Il centro del quadrato contenente sequenze con suffisso *wa* è *(x/2, y/2);*

ii. Il centro del quadrato contenente sequenze con suffisso *wc* è *(x/2, (y+1)/2);*

iii. Il centro del quadrato contenente sequenze con suffisso *wg* è *((x+1)/2, (y+1)/2);*

iv. Il centro del quadrato contenente sequenze con suffisso *wt (o wu)* è *((x+1)/2, y/2);*

f. Come conseguenza del punto 4 e 5, i pattern visibili nella rappresentazione globale CGR valgono come patterns locali. La densità (o scarsità) di punti in una regione corrisponde a un grande (o piccolo) numero di sequenze con suffissi corrispondenti alla regione. Anche, poiché ogni regione quadrata (sub-, sub-sub-, ecc. quadrante) corrisponde a un particolare suffisso, qualsiasi regione densa (o sparsa) corrisponde all'unione di S1, S2, S* in cui S_i è l'insieme di sequenze con suffisso *i*.

Per cui, volendo tracciare le informazioni che si possono trarre dalla configurazione, rientrano:

- I pattern nei CGR lungo le sequenze di DNA indicano omologia tra due sequenze.
- Il CGR consente l'identificazione di sequenze ripetitive e le loro frequenze.
- Il CGR può essere utilizzato per identificare assenti o scarse frequenze o sottosequenze.
- Le sottosequenze più brevi mostrano gli stessi modelli caratteristici come l'intera sequenza del DNA (ad esempio, i genomi).

In conclusione, utilizzare la rappresentazione CGR delle sequenze di DNA permette di distinguere facilmente diverse specie.

3 Frequency Chaos Game Representation (FCGR)

La rappresentazione Frequency Chaos Game Representation (FCGR) è una tecnica utilizzata nella bioinformatica per visualizzare e analizzare sequenze di DNA in modo non lineare. Questa metodologia trasforma sequenze di nucleotidi in immagini che rappresentano la distribuzione e la frequenza delle basi (adenina, citosina, guanina e timina) in un modo che rende visibili le proprietà frattali del DNA.

Mentre il CGR originale utilizza le coordinate esatte per ciascun punto, la discretizzazione viene chiamata rappresentazione del gioco del caos di frequenze (FCGR). Questa ha consentito un'astrazione CGR a grana grossa e meno rumorosa per sequenze. 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.

L'FCGR, simile al GR originale, consente l'identificazione di motivi o motivi mancanti in una determinata sequenza. In aggiunta, FCGR permette la visualizzazione dell'omologia tra diversi genomi a grana grossa in scala di grigi visualizzazione. I modelli in CGR sono ripetuti (cioè, i CGR sono frattali) e le sottorappresentate sequenze vengono mostrate come "buchi bianchi" nell'immagine. Inoltre, si è proposto di utilizzare FCGR per identificare elementi ripetitivi nel file genoma o duplicazioni di geni.

L'approccio utilizzato è di tipo $2^k \, x \, 2^k$, e sono basate su rappresentazioni di *k-mers.* Il numero di quadranti nella griglia di una FCGR può essere calcolato da ALMEIDA come:

$$q = 2^{2k}$$

Dove:

q è il numero di quadranti;
k è la size *k-mers*¸

Per esempio, per calcolare la lunghezza del *k-mer* di una griglia 10x10, l'equazione è:

$$k = \frac{\log_2(q)}{2}$$

Il processo di generazione di un'immagine FCGR da una sequenza di DNA segue questi passaggi:

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.

Alla fine, l'immagine risultante mostra la frequenza relativa di ogni nucleotide come zone di densità diverse. Le aree con maggiore frequenza di una particolare base saranno più evidenti.

4 Applicativi Deep learning

Le reti neurali sono modelli computazionali ispirati alla struttura e al funzionamento del cervello umano, composti da unità di elaborazione chiamate neuroni artificiali. Questi neuroni sono organizzati in strati e lavorano in sinergia per elaborare input, riconoscere pattern e generare output pertinenti. Il segreto del loro successo risiede nella capacità di apprendere da enormi quantità di dati attraverso un processo chiamato "*training*", durante il quale i parametri interni della rete vengono continuamente aggiustati per minimizzare la differenza tra l'output previsto e quello reale.

Il deep learning, una sottocategoria del machine learning, si distingue per l'utilizzo di reti neurali profonde con molti strati nascosti, permettendo alle macchine di eseguire compiti di apprendimento automatico con un livello di complessità e astrazione precedentemente irraggiungibile. Questo approccio ha dimostrato di essere particolarmente efficace nel riconoscimento di pattern complessi nei dati, dai discorsi umani alle immagini digitali, aprendo la strada a innovative applicazioni come la visione artificiale, il riconoscimento vocale e l'elaborazione del linguaggio naturale.

Le immagini costituiscono una delle forme di dati più ricche e complesse, presentando sfide uniche in termini di elaborazione e interpretazione. L'adozione del deep learning nel trattamento delle immagini ha permesso di superare tali sfide, con le reti neurali convoluzionali (CNN) che emergono come strumento di punta per l'analisi delle immagini. Queste reti sono specificamente progettate per elaborare dati strutturati a griglia, come le immagini digitali, catturando efficacemente le caratteristiche visive a vari livelli di astrazione.

4.1 Reti neurali convoluzionali

Le *reti neurali convoluzionali*, note come *CNN* (Convolutional Neural Networks), rappresentano una categoria speciale di reti neurali, ottimizzate per il riconoscimento e la classificazione di immagini. L' architettura consente alla rete di concentrarsi su piccole funzionalità di basso livello nel primo livello nascosto, quindi assemblarli in funzionalità di livello superiore più grandi nel successivo livello nascosto e così via. Questa struttura gerarchica è comune nelle immagini del mondo reale, ed è una delle ragioni per cui le CNN funzionano così bene per il riconoscimento delle immagini. [9] Le reti neurali convoluzionali utilizzano tre idee di base: *campi recettivi locali, pesi condivisi e pooling.* [8]

Negli strati connessi presenti nelle reti neurali gli input vengono rappresentati come una linea verticale di neuroni. In una rete *convoluzionale*, invece, si pensa agli input come un quadrato di neuroni di 28 x 28, i cui valori corrispondono ai pixel di 28 x 28 (esempio) *Fig. 4*:

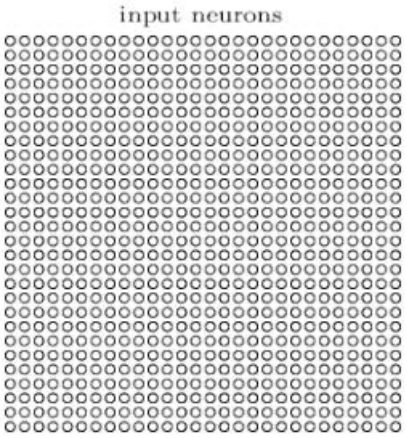


Figura 4:quadrato di neuroni input 28 x 28. .

In una rete neurale convoluzionale, l'elaborazione delle immagini non avviene attraverso un collegamento diretto e completo tra i pixel di input e ogni singolo neurone nascosto. Invece, si stabilisce un sistema di connessioni più selettivo e mirato. Ciascun neurone nel primo strato nascosto non è connesso all'intera immagine, bensì solo a una porzione ristretta di essa. *Fig.5.* Questa porzione limitata di pixel, a cui il neurone è connesso, è denominata "campo recettivo locale". Si può immaginare come una piccola finestra che scorre su tutta l'immagine di input. Ogni neurone quindi riceve in input solo i dati provenienti da questa finestra, consentendo alla rete di focalizzarsi su specifiche caratteristiche spaziali dell'immagine. Ogni connessione apprende un *peso* e il neurone nascosto apprende un comportamento generale.

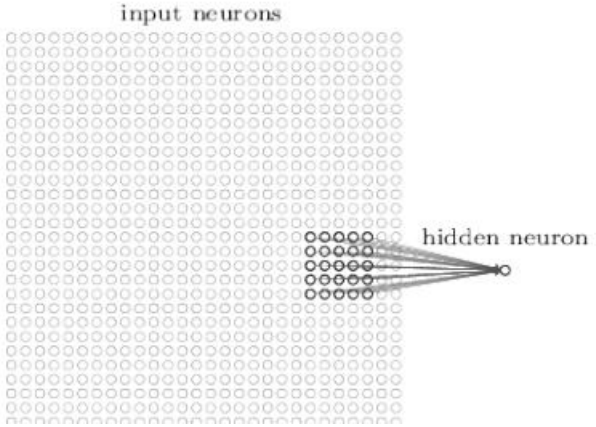


Figura 5: processo di convoluzione.

Si può concepire ogni neurone del primo strato nascosto di una rete neurale convoluzionale come specializzato nell'analisi di una specifica sezione dell'immagine di input, nota come *campo recettivo locale* . Per esempio, immaginiamo di iniziare con un campo recettivo che copre l'angolo superiore sinistro dell'immagine: questa sarà l'area di competenza del primo neurone nascosto. Poi, spostiamo questo campo recettivo di un pixel verso destra, e questo

nuovo insieme di pixel diventerà l'input per il secondo neurone nascosto. In questa maniera, ogni neurone nascosto nel primo strato "guarda" a una differente area dell'immagine, apprendendo dalle particolarità di quella specifica sezione.

Man mano che il campo recettivo si muove sull'intera immagine, un mosaico di neuroni nel primo strato nascosto lavora per identificare e imparare da varie parti dell'immagine, permettendo alla rete di costruire una comprensione dettagliata e frammentata dell'input visivo, *Figura 6*. E così via, si sta costruendo il primo strato nascosto. Più in generale, ogni neurone locale nella riga i, colonna j di un dato layer è connesso all output del neurone del layer precedente posizionato nella riga i a $i + f_h - 1$, colonna j a $j + f_w - 1$, dove f_h *e* f_w sono i pesi e le altezze del campo recettivo. Per fare in modo che uno strato abbia la stessa altezza e larghezza del layer precedente è normale aggiungere degli zero intorno agl input, questo processo è definito *zero padding.*

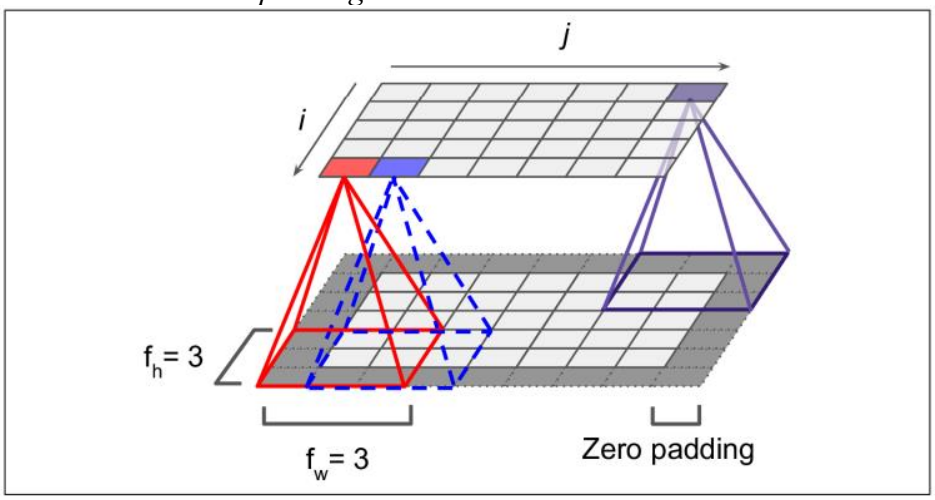


Figura 6: Processo di convoluzione e padding.

Ogni neurone nascosto ha un *bias* e n x n pesi legati al suo campo recettivo locale. Ciascun peso e bias viene utilizzato per ciascuno dei neuroni nascosti. In altre parole, per j, *k-esimo* neurone nascosto, l'output è :

$\sigma\left(b + \sum_{l=0}^{4} \sum_{m=0}^{4} w_{l,m} a_{j+l,k+m}\right)$

Qui, σ è la funzione di attivazione, b è il valore condiviso per il bias. $w_{l,m}$è un array n x n di pesi condivisi, $a_{j+l,k+m}$ denota l'input di attivazione alla posizione *x,y.* Questo significa che tutti i neuroni nel primo layer nascosto rilevano esattamente la stessa caratteristica, solo in posizioni diverse dell immagine. Per le reti convoluzionali vale il principio di invarianza su immagine: spostare di poco l'immagine di un gatto rimarrà sempre l'immagine di un gatto.

I pesi condivisi e il bias definiscono un *kernel* o un filtro. Il peso di un neurone può essere rappresentato attraverso una piccola immagine della taglia del campo recettivo, mentre ogni filtro è una matrice 2D che ha la dimensione del campo recettivo locale. Questa matrice viene fatta scorrere sull'immagine di input, e in ogni posizione, i valori dei pixel dell immagine vengono moltiplicati per i corrispondenti pesi del filtro e poi sommati insieme, aggiungendo il bias al risultato. Un filtro, quindi, è in grado di evidenziare specifiche caratteristiche nell'immagine, come bordi o angoli, a seconda di come i suoi pesi sono stati addestrati durante il processo di apprendimento. Quando il filtro si muove sull'intera immagine, produce una nuova immagine 2D chiamata "*feature map*" o

mappa delle caratteristiche, che rappresenta la presenza e l'intensità di una specifica caratteristica in diverse posizioni dell'immagine. Un livello convoluzionale è formato da più di questi filtri, ognuno dei quali è addestrato a riconoscere un tipo differente di caratteristica dall'immagine di input. Ogni filtro produce una *sua feature map* unica; l'insieme di tutte queste feature maps costituisce l'output completo del livello convoluzionale, fornendo una rappresentazione multidimensionale delle varie caratteristiche rilevate nell'immagine di input, *Fig. 7.*

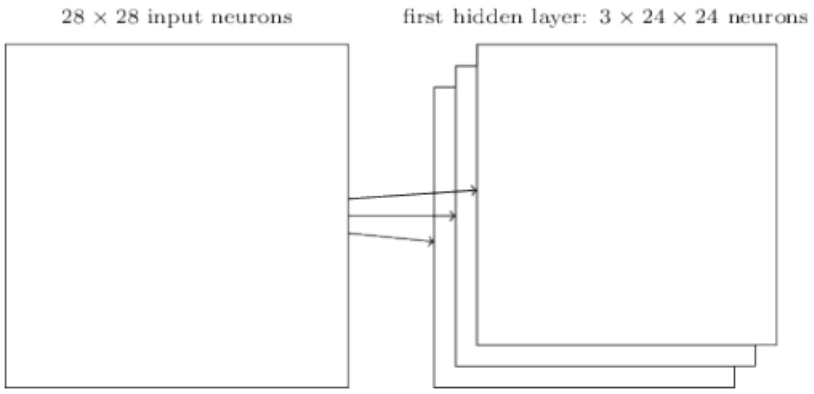


Figura 7: Processo di acquisizione di più livelli.

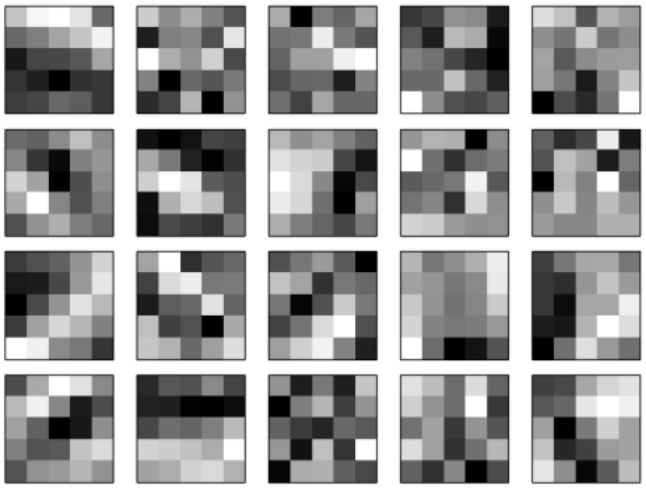


Figura 8: esempi di features map 5 x 5.

Le immagini corrispondono, ad esempio, a 20 differenti *feature map*s (kernels), *Fig. 8*. Ogni mappa è rappresentata come un'immagine a blocchi 5×5, corrispondente ai pesi 5×5 presenti nel campo recettivo locale. I blocchi più

bianchi indicano un peso più piccolo (tipicamente più negativo), quindi la mappa delle caratteristiche

risponde meno ai pixel di input corrispondenti. I blocchi più scuri significano un peso maggiore, quindi la mappa delle caratteristiche risponde maggiormente ai pixel di input corrispondenti.

Ancora, la figura successiva mostra due possibili insiemi di pesi. Il primo è rappresentato come un quadrato nero con una linea bianca verticale all'interno

al centro (è una matrice 7 × 7 piena di 0 ad eccezione della colonna centrale, che è piena di 1); i neuroni che utilizzano questi pesi ignoreranno tutto nel loro campo recettivo tranne per la linea verticale centrale (poiché tutti gli input verranno moltiplicati per 0, ad eccezione di quelli situati nella linea verticale centrale). Il secondo filtro è un quadrato nero con a linea bianca orizzontale al centro. Ancora una volta, i neuroni che utilizzano questi pesi lo faranno ignorano tutto nel loro campo recettivo tranne la linea orizzontale centrale.

Ora, se tutti i neuroni in uno strato utilizzano lo stesso filtro a linee verticali (e lo stesso termine di bias), e si fornisce alla rete l'immagine di input mostrata nella *Fig. 9* (immagine in basso), il layer produrrà l'immagine in alto a sinistra. Allo stesso modo, l'immagine in alto a destra è ciò che si ottiene se tutti i neuroni utilizzano lo stesso filtro di linea orizzontale. Un intero strato di neuroni applica lo stesso filtro o kernel all'immagine di input per creare una mappa delle caratteristiche. Questa mappa evidenzia specifiche parti dell'immagine che rispondono fortemente a quel filtro, indicando la presenza di particolari tratti o pattern che il filtro è stato progettato per catturare. Non è necessario definire questi filtri a mano; piuttosto, sono i filtri stessi ad essere appresi dalla rete durante il processo di allenamento. Attraverso l'addestramento con un vasto set di dati, la rete impara automaticamente quali caratteristiche sono importanti per compiere le task di riconoscimento o classificazione delle immagini.

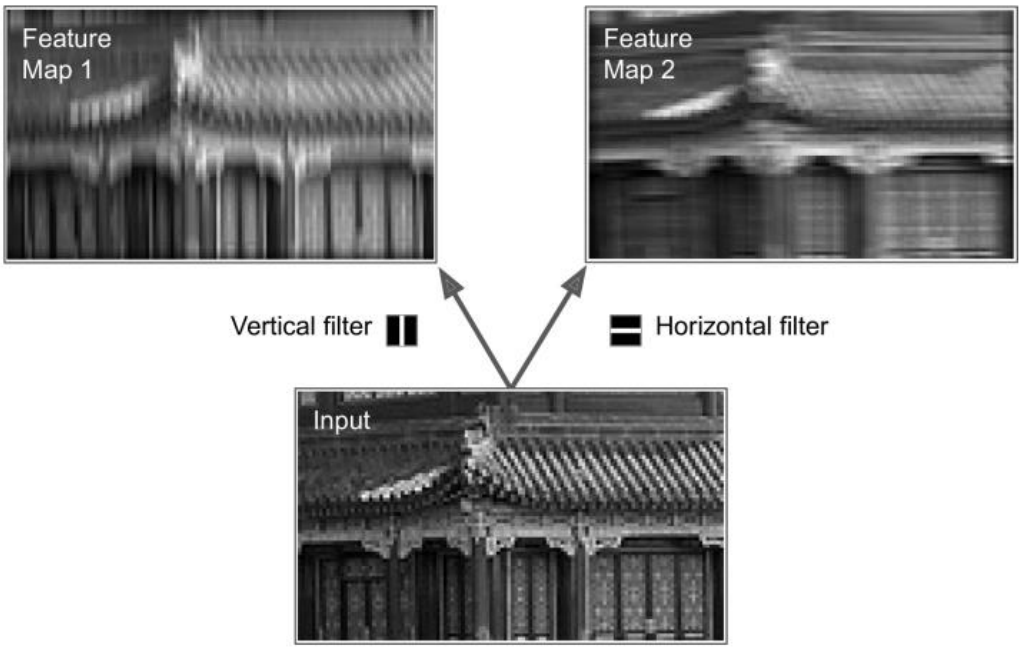


Figura 9: Esempio di applicazione di filtri verticali e orizzontali.

Le reti neurali convoluzionali includono strati di pooling, il cui scopo è ridurre la dimensione delle informazioni generate dagli strati convoluzionali. In pratica, uno strato di pooling riceve come input le mappe delle caratteristiche prodotte dallo strato convoluzionale e produce una versione semplificata di queste mappe. Un metodo comune utilizzato in questi strati è il max-pooling: in questa operazione, ogni cella di pooling seleziona il valore massimo di attivazione da una regione specifica della mappa delle caratteristiche. Il max-pooling viene eseguito su ogni mappa delle caratteristiche in modo indipendente, risultando in un insieme di mappe ridotte che conservano le informazioni più salienti ma con una risoluzione minore. Questo processo non solo aiuta a rendere la rete più efficiente, riducendo il numero di parametri da elaborare nei livelli successivi, ma contribuisce anche a rendere la rete più robusta a piccole variazioni e distorsioni nell'immagine di input.

Il *max-pooling* può essere inteso come il meccanismo attraverso il quale la rete neurale identifica la presenza di una caratteristica specifica entro una certa area dell'immagine, indipendentemente dalla sua posizione precisa. L'idea di fondo è che, una volta rilevata una caratteristica, la sua localizzazione esatta è meno critica di una sua localizzazione relativa; ciò che conta è che la caratteristica sia stata rilevata all'interno di una determinata zona. Questo principio aiuta la rete a comprendere e a identificare le caratteristiche visive in maniera più astratta e invariante rispetto alle trasformazioni geometriche come traslazioni o distorsioni lievi.

Avendo definito i principi cardine, è possibile sintetizzarli per costruire un'architettura di rete neurale convoluzionale funzionante. Al termine degli strati convoluzionali e di pooling, si trova il livello completamente connesso, o fully-connected layer. Questo strato crea una connessione densa, collegando ogni neurone proveniente dallo strato di pooling precedente a ogni neurone nel livello di output. Il compito di questo strato è di integrare le caratteristiche rilevate, che sono state astratte e ridotte nei livelli precedenti, per formare la predizione finale o la classificazione.

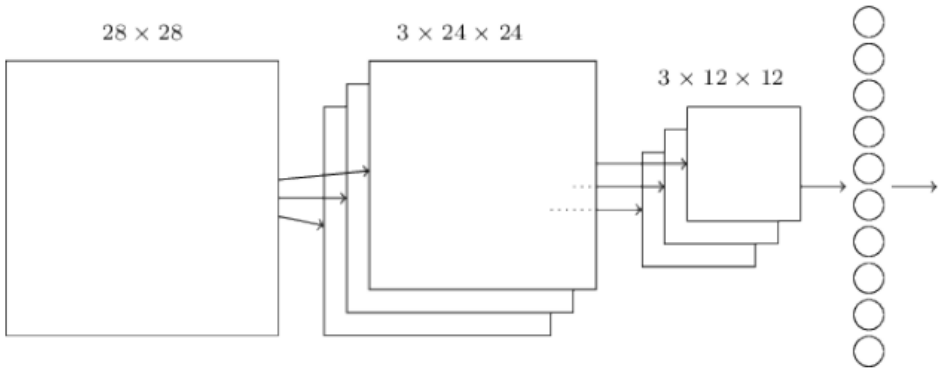


Figura 10: Processo completo.

In conclusione, una rete neurale è costituita da una serie di unità di base, ovvero i neuroni, i cui comportamenti sono regolati dai relativi pesi e bias. L'obiettivo fondamentale nell'addestramento di una rete neurale è ottimizzare questi pesi e bias utilizzando un set di dati di addestramento. Attraverso questo processo, si

affina la capacità della rete di eseguire correttamente classificazioni o altre operazioni sui dati di input. L'addestramento è una fase critica durante la quale la rete apprende come trasformare l'input in output desiderato, adattando i suoi parametri interni (pesi e bias) per ridurre l'errore tra le previsioni della rete e le etichette reali fornite nei dati di addestramento.

CAPITOLO 3

PRESENTAZIONE DEI DATI E DISCUSSIONE DEI RISULTATI

1 Analisi dei dati clinici

1.1 Descrizione del dataset

I dati relativi alle sequenze genomiche e ai metadati associati sono stati resi disponibili pubblicamente attraverso il database EpiCov di GISAID. Questo database comprende circa 9 milioni di sequenze genomiche individuali, raccolte da 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.

I dataset utilizzati, "*data clinical patient*" e "*data covid patient*" si compongono di 369181 osservazioni (pazienti) su 28 variabili. Questi riportano diverse informazioni genetiche e mediche relative ai pazienti rilevati per lo studio in questione. Di seguito vengono riportate le variabili di cui si compongono i dataset.

Tabella 1: Descrizione delle variabili *data clinical patient*

Variabile	Descrizione
1. ACCESSION ID	Un identificatore univoco assegnato a ciascuna sequenza di virus. Viene utilizzato per tracciare e riferire specifiche sequenze genomiche nei database.
2. TYPE	Specifica il tipo di virus; in questo caso, tutte le sequenze sono indicate come "betacoronavirus", il genere a cui appartiene SARS-CoV-2.
3. CLADE	Indica il clade virale, che è una suddivisione del virus basata su mutazioni genetiche che definiscono diversi gruppi (clade). "G" è uno dei cladi principali.
4. PANGO LINEAGE	Il lignaggio del virus secondo la classificazione Pango, un sistema che categorizza le varianti di SARS-CoV-2 basandosi sulle loro caratteristiche genetiche. Esempi includono B.1, P.1.17, etc.
5. PANGO VERSION	La versione del sistema di classificazione Pango utilizzata per assegnare il lignaggio.
6. AA SUBSTITUIONS	Un elenco di sostituzioni di amminoacidi nelle proteine virali. Queste mutazioni possono influenzare la trasmissibilità del virus, la virulenza e la capacità di evadere la risposta immunitaria.
7. VARIANT	Fornisce il nome della variante, se identificata, come "Gamma" o "Eta", e una breve descrizione della sua origine o delle sue caratteristiche principali.

8. PASSAGE HISTORY	Informazioni sul passaggio del virus attraverso ospiti o colture cellulari, che può influire sul suo genoma.
9. COLLECTION DATE	La data in cui il campione è stato raccolto.
10. LOCATION	La località geografica di raccolta del campione, specificata fino al livello della città o regione.
11. HOST	L'ospite dal quale è stato prelevato il campione, tutti umani.
12. ADDITIONAL LOCATION INFORMATION	Informazioni addizionali sulla condizione del paziente.
13. GENDER	Sesso biologico del paziente.
14. AGE	Età del paziente.
15. STATUS	Stato di sopravvivenza del paziente. Questa variabile assume i seguqnti livelli: prevalentemente sconosciuto, vivo, ospedalizzato, rilasciato, sintomatico.
16. SPECIMEN SOURCE	Tipo di campione biologico prelevato dai pazienti per l'analisi del virus. Questo campione viene utilizzato per eseguire test diagnostici, come il test PCR (Polymerase Chain Reaction) o altri tipi di analisi genomiche, per determinare la presenza del virus nel corpo del paziente. Ad esempio : tampone nasofaringeo, tampone orofaringeo, tampone nasale
17. ADDITIONAL HOST INFORMATION	Stato somatico
18. SAMPLING STRATEGY	Il metodo utilizzato per raccogliere campioni biologici dai pazienti. Maggiori dettagli sulle strategie di campionamento : 1) **Random sampling:** Viene selezionato un campione rappresentativo del paziente generale senza particolari criteri di selezione, garantendo che ogni campione della popolazione abbia la stessa probabilità di essere selezionato. **2) Targeted sampling:** Viene effettuato un campionamento in gruppi specifici di relazioni, nell ambito di pazienti con sintomi specifici, o individui di aree geografiche particolari.
19. LAST VACCINATED	Informa se il soggetto da cui è stato raccolto il campione era stato vaccinato recentemente (se disponibile).
20. TREATMENT	Informazioni riguardanti eventuali trattamenti ricevuti dal paziente, se disponibili.
21. SUBMISSION DATE	La data in cui la sequenza è stata caricata nel database, utile per capire quando la sequenza è diventata disponibile alla comunità scientifica.

22. IS REFERENCE?	Indica se la sequenza è utilizzata come riferimento per altre sequenze, cioè se rappresenta un esempio "standard" del virus.
23. IS COMPLETE?	Specifica se la sequenza è completa o parziale.
24. IS HIGH COVERAGE?	Indica se la copertura della sequenza è alta, cioè se la sequenza è stata determinata con alta precisione.
25. IS LOW COVERAGE?	Indica se la copertura della sequenza è bassa, segnalando possibili incertezze nella determinazione della sequenza.
26. N-CONTENT	Proporzione di nucleotidi 'N' nella sequenza, che rappresentano basi non determinate. Un alto contenuto di 'N' può indicare una sequenza di qualità inferiore.
27. GC-CONTENT	Percentuale di guanina (G) e citosina (C) nella sequenza, una misura comune per valutare la composizione del genoma.
28. SEQUENCE LENGTH	Lunghezza della sequenza in nucleotidi. Le variazioni nella lunghezza possono indicare sequenze incomplete o la presenza di inserzioni o delezioni.

1.2 Data Cleaning e Data Wrangling

Il “data wrangling,” chiamato anche “data munging” rappresenta un passaggio cruciale nell’analisi dei dati. Molto spesso si hanno a disposizione un insieme di dati grezzi provenienti da diverse fonti. Questi dati possono essere disorganizzati, contenere errori, dati mancanti o informazioni in formati diversi. Senza una preparazione adeguata, l’analisi e la modellazione dei dati sarebbero difficili, se non impossibili. Il data wrangling è il processo di trasformazione dei dati disordinati e sporchi in un formato coerente e adatto all’analisi. Il passaggio comporta molteplici attività, tra cui la pulizia dei dati per correggere errori e rimuovere duplicati, la standardizzazione delle unità di misura, la trasformazione dei dati categorici in forme numeriche comprensibili, e la creazione di nuove variabili o caratteristiche quando necessario. L’obiettivo finale del data wrangling è creare un dataset pulito, coerente e pronto per essere analizzato, riducendo così i potenziali errori e garantendo che i risultati dell’analisi siano accurati e significativi.

I dati grezzi sono spesso *disordinati e formattati male.* Inoltre, potrebbero mancare definizioni appropriate che tengano conto della scala di misurazione utilizzata.

Per cui la pulizia dei dati consiste nel procedimento mediante il quale si esaminano e si migliorano i dati contenuti nel dataset, con l'obiettivo di assicurare che siano di alta qualità e validi per l'analisi statistica. Le procedure che caratterizzano questo passaggio sono le seguenti:

- **Analisi dei missing values.** Si verifica la presenza/assenza dei valori mancanti (NA) che può notevolmente influire sull'analisi. Si procede , quindi, decidendo di eliminarli o sostituendoli con valori reali (esempio: imputazione di media o mediana).

- **Individuazione di valori anomali.** Si esamina attentamente il dataset per individuare eventuali valori inesatti o insoliti, e si prendono decisioni su come trattarli. Queste decisioni possono includere l'eliminazione dei valori problematici o la loro sostituzione con valori appropriati.
- **Trasformazione delle variabili.** Si verifica che tutte le variabili abbiano la classe appropriata *Double o integer* per i numeri, *factor* per le variabili categoriali.
- **Implementazione di nuove variabili.** Sulla base delle variabili già presenti nel dataset si possono creare delle nuove variabili, ad esempio facendo operazioni matematiche tra due variabili o creando una variabile multilivello sulla base di una variabile numerica.

a) Analisi dei missing values

Per il dataset, *tabella 2,* sono mancanti un numero abbastanza elevato di osservazioni , con principale riferimento alle variabili *VARIANT, ADDITIONAL LOCATION INFORMATION, LAST VACCINATED E TREATMENT.* Poiché queste variabili non è possibile costruirle a partire da altre informazioni e tenendo conto che verranno utilizzate per ottenere informazioni descrittive tramite metodologie in grado di lavorare con dati censurati, si decide di mantenere tutte le osservazioni.

b) Individuazione di valori anomali

L'individuazione di possibili valori anomali avviene attraverso l'esplorazione del dataset.

Nello studio dei dati, si osserva che ci sono 75480 individui identificati come femmine e i restanti come maschi. Inoltre, il ignaggio pango più frequente è B.1, non assegnato e B.1.1.7. E' inoltre presente un errore di trascrizione nella variabile relativa alla storia del paziente. Al momento della registrazione dei dati circa 14000 pazienti provengono dall'inghilterra, circa 11000 dal Nord America, la stessa cifra dalla turchia. Relativamente alla location del campione è necessario sottolineare che i dati sono stati aggregati da fonti diverse per cui è probabile che la trascrizione possa essere avvenuta senza tener conto di una linea standard. Per alcuni pazienti la registrazione è avvenuta rispetto al paese di origine, per altri è stato segnalato anche la città di appartenenza. Lo stesso vale per il tipo di campionamento, potrebbe essere lo stesso ma segnato con nomenclatura diversa. La percentuale più bassa di GC è del 34% mentre la più alta del 45%. Circa il 50% dei pazienti ha una percentuale di nucleotidi oscillante tra 0.03 e 0.07 per cento, il valore minimo osservato è 0 e il massimo è 0.99. Poiché media e terzo quartile coincidono è dubbia la presenza del 25% dei pazienti che ha valori compresi tra 0.07 e 99.

Tabella 2: Summary *data clinical patient*

Accession ID	Type	Clade	Pango lineage	Pango version
Length:369181	betacoronavirus:369181	G :369173	B.1 : 73114	consensus call :279632
Class :character		NA's: 8	Unassigned: 68724	PANGO-v1.23 : 86722
Mode :character			B.1.1.7 : 22271	SCORPIO_v0.1.12: 2334
			B.1.243 : 17882	PANGO-v1.23.1 : 459
			B.1.258 : 17705	PANGO-v1.22 : 12
			(Other) :169477	(Other) : 14
			NA's : 8	NA's : 8

```
                Passage details/history Collection date                                  Location
Original                       :360778   Length:369181   Europe / United Kingdom / England    : 14176
Orginal                        :  6844   Class :character North America / USA / Texas / Houston: 11287
unknown                        :   202   Mode  :character Europe / Turkey                     : 11231
Passage 1, Covid-test, NS500:    181                     Europe / Austria / Vienna            :  8015
Original isolate               :   173                   Asia / South Korea                   :  6930
P1, Covid-test, NS500          :   158                   Europe / Austria / Tirol             :  6326
(Other)                        :   845                   (Other)                              :311216
```

```
     Host        Additional location information     Gender       Patient age
Length:369181   Length:369181                    unknown:221681   Length:369181
Class :character Class :character                Male   : 75480   Class :character
Mode  :character Mode  :character                Female : 71547   Mode  :character
```

```
                                Sampling strategy Last vaccinated                Treatment
Baseline surveillance                   : 24120   Length:369181   -                    :   353
Random sampling                         :  5204   Class :character Illumina NextSeq    :   222
Baseline surveillance (random sampling):  2875   Mode  :character Nanopore MinION     :   200
Suspect sampling                        :  2140                   No current treatment:   145
Baseline Surveillance                   :  1811                   standard of care     :    25
(Other)                                 : 12722                   (Other)              :   237
NA's                                    :320309                   NA's                 :367999
```

```
 Submission date     Is reference?  Is complete?  Is high coverage? Is low coverage?  N-Content
 Min.   :2020-01-31  Mode:logical   Mode:logical  Mode:logical      Mode:logical      Min.   :0.00
 1st Qu.:2021-02-27                TRUE:308478   TRUE:192450       TRUE:102667       1st Qu.:0.01
 Median :2021-05-15                                                NA's:266514       Median :0.03
 Mean   :2021-06-26                                                                  Mean   :0.07
 3rd Qu.:2021-09-08                                                                  3rd Qu.:0.07
 Max.   :2024-07-12                                                                  Max.   :0.99
                                                                                     NA's   :118167
  GC-Content      Sequence length
 Min.   :0.3442   Min.   :  142
 1st Qu.:0.3793   1st Qu.:29614
 Median :0.3798   Median :29782
 Mean   :0.3780   Mean   :25760
 3rd Qu.:0.3801   3rd Qu.:29823
 Max.   :0.4533   Max.   :43996
```

1.3 Statistical Analysis

Nell'ambito dell'analisi è stato condotto uno studio per esaminare le associazioni tra diverse variabili di interesse. Per valutare il legame e il grado di associazione si ricorre al test del chi quadro, i cui principi di funzionamento sono illustrati nel Paragrafo 1.1 del Capitolo 2.

Tramite il test del chi quadrato di Pearson, si intende controllare se l'associazione fra due variabili sia statisticamente significativa. Il test del chi quadrato confronta i valori osservati di frequenza in una tabella di contingenza con i valori attesi, che rappresentano l'ipotesi di indipendenza tra le variabili, l'ipotesi nulla (H_0) del test afferma che le due variabili sono indipendenti l'una dall'altra, mentre l'ipotesi alternativa (H_1) afferma che le due variabili sono associate in qualche modo. Le caratteristiche demografiche, di laboratorio e cliniche sono state confrontate sulla base della variante identificata, tenendo conto dei risultati del test. Sono state selezionate le sole 4 varianti più frequenti, nello specifico:

1) Variante VOC Alpha, identificata per la prima volta nel Regno Unito;

2) Variante VOI Eta, identificata per la prima volta nel Regno Unito e in Nigeria;
3) Variante VOC Omicron, identificata per la prima volta in Botswana, Hong Kong e South Africa;
4) Variante VOC Gamma, identificata per la prima volta in Brasile e Giappone.

Per il solo scopo di analisi, poiché le categorie sono molte, vengono riportate solo quelle con maggiore frequenza. Oltretutto, per il solo scopo esplorativo, la variabile numerica che riguarda l'età dei pazienti è stata categorizzata in fasce da 0-10 (bambini), 11-28 anni (adolescenti e ragazzi), 29-50(adulti), 50+ (anziani).

Stabilito il valore di α=0.05,

La probabilità che l'associazione sia dovuta al caso è prossima allo zero per lignaggio pango, pango version, genere, status del paziente e età. L'ipotesi nulla viene respinta e l'associazione è statisticamente significativa. Diversamente per la variabile dello specimen type , il valore del p-value porta a non rifiutare l'ipotesi nulla stabilendo l'indipendenza fra le coppie di variabili.

Tabella 3: Risultati del test Chi-Quadrato

Variabile	**VARIANT ALPHA**	**VARIANT ETA**	**VARIANT OMICRON**	**VARIANT GAMMA**	**P-value**
Pango version					<< 2.2e-16***
Consensus call	21559 (95%)	9526 (95%)	8414 (98%)	3193 (49%)	
Pango v1.23	422 (2%)	403 (4%)	14 (1%)	3208(49%)	
Scorpio 1.12	598 (3%)	62 (1%)	30 (1%)	88(2%)	
Gender					
Male	50165 (51%)	50035 (56%)	45487 (49%)	49125 (53%)	0.03 *
Female	40809 (49%)	40028 (45%)	46396 (51%)	41804 (47%)	
Patient status					<< 2.2e-16***
Live	142 (22%)	525 (92%)	161 (80%)	125 (96%)	
Released	504 (78%)	44 (8%)	40 (20%)	5 (4%)	
Età					0.0001***
0 -10	528 (5%)	204 (4%)	172 (3%)	180 (4%)	
11-28	2294 (25%)	1072 (25%)	1069 (23%)	1039 (24%)	
29-50	3298 (35%)	1820 (41%)	1907 (41%)	1656 (38%)	
50+	3303 (35%)	1250 (28%)	1440 (31%)	1376 (32%)	

PRIMARY					
Specimen source			796 (35%)	1492 (88 %)	
		901 (49 %)	1426 (63	187 (10%)	
NASOPHARYNGEAL SWAB	2385(52 %)	727 (45 %)	%)	11(1 %)	
	2195 (42 %)	266 (5 %)	33 (2 %)		0.07
	239 (6 %)				
OROPHARYNGEAL SWAB					
NASAL SWAB					
ALTRO					

La *Tabella 3* riporta il valore associato del p-value per ciascuna categoria oltre che la partizione utile per descrivere le caratteristiche principali della popolazione. 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, la differenza è statisticamente significativa anche se non in maniera molto forte (p-value = 0.03).

La prevalenza di clade O nell'intera coorte è del 71%, *Fig. 11 lettera a*. Dall'analisi comparativa delle varianti in relazione ai siti di prelievo del campione, *Fig. 11 lettera c*, è emerso che la maggior parte dei tamponi sono nasofaringiali, a seguire orofaringali e infine. Non è presente nessuna associazione statisticamente significativa tra le varianti e questa variabile (p-value >0.05). Dall'analisi comparativa delle varianti in relazione allo status dei pazienti, *Fig.11 lettera d*, è emerso che , in maniera prevalente, per le varianti omicron, gamma e eta i pazienti sono sopravvisuti e le ospedalizzazioni sono pressocchè minime. Tuttavia, per la variante alpha, una forte percentuale è stata ospedalizzata e poi rilasciata (78% vs 8% vs 20% vs 4%), probabilmente è un tipo di variante che ha dei sintomi più gravi sui pazienti, con una significativa differenza statistica ($p < 2.2e-16$). Ancora, *Fig.11 lettera e*, la barra associata alla variante Alpha è nettamente la più alta, indicando che questa variante ha colpito un numero significativamente maggiore di individui rispetto alle altre varianti considerate. Per cui, la diffusione della variante Alpha è stata più ampia, indipendentemente dall'età. Il grafico evidenzia che per tutte le varianti considerate, la fascia d'età 50+ anni è quella più colpita. La predominanza di questa fascia d'età è particolarmente evidente per la variante Alpha, ma è presente anche nelle altre varianti. Un'altra osservazione significativa è la distribuzione relativamente uniforme dei casi tra le altre fasce d'età (0-10 anni, 11-28 anni, 29-50 anni) per le varianti Omicron e Gamma. Il p-value riportato nel grafico, pari a 0.00001, è particolarmente significativo e indica una forte associazione statistica tra le varianti e la distribuzione dei casi tra le diverse fasce d'età.

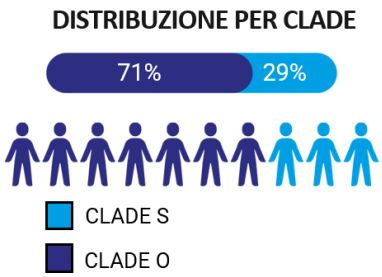
La nomenclatura Pango , utilizzata per distinguere le varianti di malattia e monitorare le diffusioni dell'epidemia sono così distinte, *Fig. 11 lettera f*:

1. **B.1.1.7**, rappresenta il 47% del totale delle osservazioni. Non sono state trovate evidenze in letteratura che le persone con Covid-19 B.1.1.7., la cosiddetta variante inglese, presentino sintomi peggiori o un rischio maggiore di sviluppare long Covid rispetto a quelli infettati con un diverso ceppo. Tuttavia, la carica virale e il numero R erano più alti per B.1.1.7., confermando l'evidenza che la variante inglese è più trasmissibile del ceppo originario rilevato a Wuhan. [10]
2. **B.1.525**, che costituisce il 21% delle osservazioni. Il lignaggio B.1.525 è associabile alla variante eta. Il lignaggio B.1.525 condivide alcune mutazioni con altre varianti preoccupanti, come la variante Alpha (B.1.1.7). Tra queste mutazioni, alcune delle più rilevanti sono: E484K: Questa mutazione nella proteina Spike è stata associata a una potenziale riduzione dell'efficacia degli anticorpi, sia quelli derivanti da infezione naturale sia quelli indotti dai vaccini. Q677H: Mutazione nella proteina Spike, che potrebbe influenzare la trasmissibilità del virus. F888L: Un'altra mutazione nella proteina Spike, la cui rilevanza è ancora oggetto di studio.
3. il lignaggio **P.1**, costituisce il 12% del campione, è assimilabile alla variante Gamma. La variante P.1 è caratterizzata da diverse mutazioni nella proteina Spike, che è la proteina che il virus utilizza per entrare nelle cellule umane. Alcune delle mutazioni più rilevanti includono: E484K: Questa mutazione è associata a una riduzione della neutralizzazione da parte degli anticorpi, sia quelli derivanti dall'infezione naturale che dalla vaccinazione. K417T: Questa mutazione nella proteina Spike è simile a quella osservata nella variante Beta (B.1.351) ed è stata associata a un aumento della capacità del virus di legarsi al recettore ACE2 umano. N501Y: Questa mutazione è presente anche nelle varianti Alpha (B.1.1.7) e Beta (B.1.351), ed è associata a un aumento della trasmissibilità del virus.

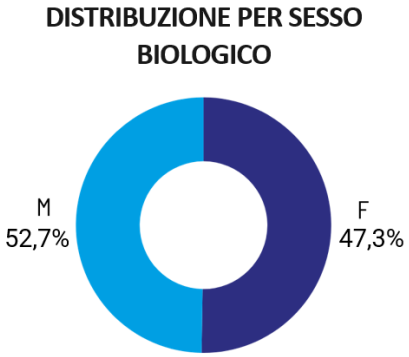
Figura 11: Overview della corte. (A) Distribuzione con infografica della clade O e S; (B) Distribuzione piechart dei pazienti per sesso biologico; (C) Distribuzione della sitologia di prelievo del campione; (D) Stacked bar chart della distribuzione di status di

sopravvivenza in relazione alle varianti; (E) Stacked bar chart della distribuzione di età rispetto alle varianti; (F) Distribuzione della pango lineage.

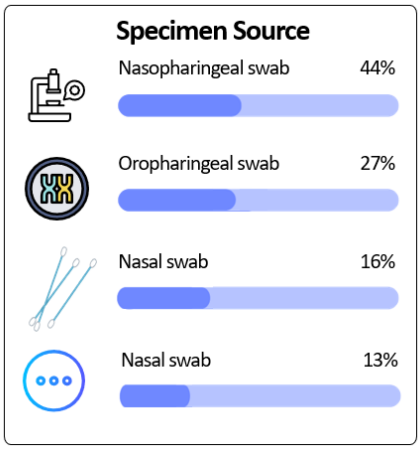
A)



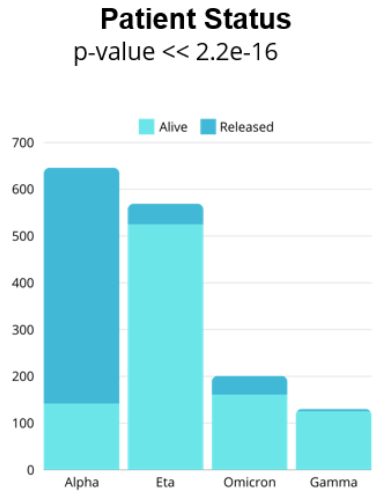
B)



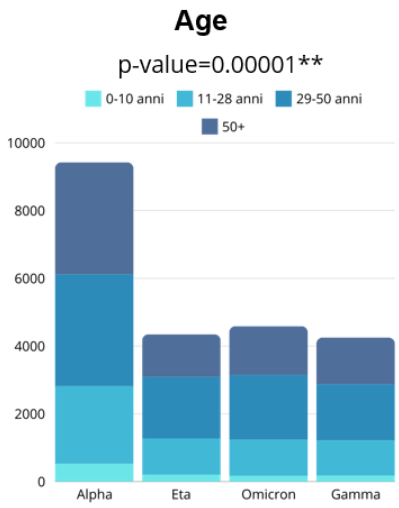
C)



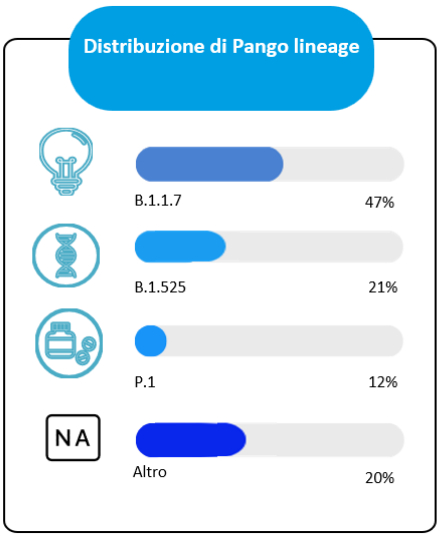
D)



E)



F)



2 EDA

Nell'ambito dell'analisi è stato condotto un'analisi esplorativa per esaminare le associazioni tra diverse variabili di interesse, oltre che prevalenze di classe. Il *dataset* contiene informazioni sulle sostituzioni degli amminoacidi, sono racchiuse nella colonne "*AA subsistion"* (è stato necessario applicare delle tecniche di estrazione per rendere facilmente visualizzabile il risultato). Le sostituzioni di amminoacidi indicano cambiamenti specifici nelle proteine del virus, dove un amminoacido è stato sostituito da un altro rispetto alla sequenza di riferimento originale del virus. Questo tipo di mutazione può avere effetti diversi, che vanno da nessun impatto a significative alterazioni della funzionalità delle proteine, che a loro volta possono influenzare la trasmissibilità, la virulenza, o l'evasione immunitaria del virus.

Alcune proteine *(Fig. 13.A)*, come le proteine NSP3 o NSP12, presentano un numero relativamente elevato di mutazioni, con Spike che mostra il conteggio più alto tra tutti con oltre 1.75 e^6 mutazioni. In contrasto, altre proteine, come NSP1, NSP14 e NSP13, presentano un numero molto più basso di mutazioni (per alcune prossimo all'1).

Il barplot in *Fig. 13.B*, 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, una componente cruciale del virus per l'infezione delle cellule umane.

La matrice di co-occorenza in *Fig. 13.C* 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. Inoltre, Spike_H69del appare in diversi campioni insieme a una varietà di altre mutazioni, come indicato dalle celle blu chiaro nella sua riga.

C'è una notevole variabilità nel numero di mutazioni tra i campioni nella *Fig. 13.D*. Questo riflette l'eterogeneità tumorale, esprimendo che ogni tumore ha un profilo genetico unico che potrebbe influenzare la risposta al trattamento e l'outcome clinico.

2.1 Analisi delle proteine e delle mutazioni

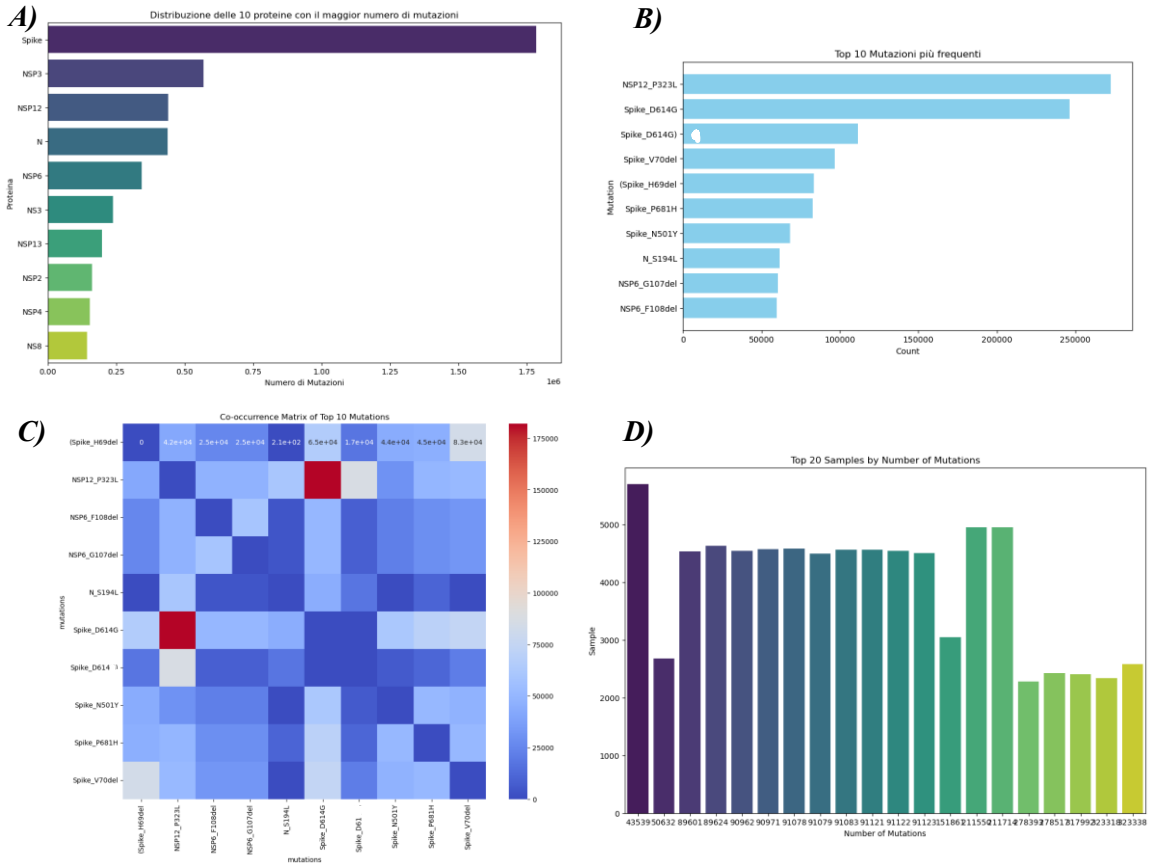
E' stata condotta un'analisi dettagliata delle mutazioni genetiche presenti nei dati e delle principali proteine responsabili.

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 della proteina spike di SARS-CoV-2 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 ACE2 sulla cellula bersaglio, 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. Infatti, bloccarne il funzionamento vorrebbe dire impedire al virus di infettare le cellule bersaglio, rendendolo quindi innocuo.[16] Tra le mutazioni più frequenti è presente la D614G: secondo [17] i pazienti infettati con il picco che esprime SARS-CoV-2 (virus G614) hanno prodotto titoli infettivi più elevati nei lavaggi nasali e nella trachea, ma non nei polmoni, supportando prove cliniche che dimostrano che la mutazione aumenta la carica virale nel tratto respiratorio superiore ai Pazienti affetti da COVID-19 e possono aumentare la trasmissione. La seconda mutazione più frequente è V70del, quest'ultima è identificata da [18] come una delle mutazioni più frequenti trasmesse dai visoni agli esseri umani e portatrice di possibili conseguenze gravi sui pazienti. Ancora, la mutazione P681H. Prove sempre più numerose suggeriscono che le varianti preoccupanti (VOC) di SARS-CoV-2 si evolvono per eludere la risposta immunitaria umana, con grande interesse concentrato sulle mutazioni nella proteina spike che sfugge agli anticorpi. In [19] viene dimostrato che il COV alfa (B.1.1.7) è sostanzialmente più resistente agli interferoni di tipo I rispetto al virus parentale simile a Wuhan. Ciò è correlato alla resistenza alla proteina antivirale IFITM2 e al potenziamento da parte del suo paralogo IFITM3. Il determinante chiave di ciò è un cambiamento da prolina a istidina nella posizione 681 in S adiacente al sito di clivaggio della furina.

NSP3: NSP13 è una proteina da 67 kDa che appartiene alla superfamiglia dell'elicasi 1B. Utilizza l'energia dell'idrolisi del nucleotide trifosfato per catalizzare lo svolgimento del DNA o dell'RNA a doppio filamento in una direzione da 5′ a 3′7. Sebbene si ritenga che NSP13 agisca sull'RNA in vivo, la caratterizzazione enzimatica mostra un'attività significativamente più robusta sul DNA nei test in vitro con un'attività dell'elicasi non processiva relativamente debole rispetto ad altri enzimi della superfamiglia 1B8,9. [20] Nsp13 è essenziale per la replicazione e la propagazione di tutti i coronavirus umani e non umani. In combinazione con il suo sito di legame nucleotidico definito e la sua farmacoponibilità, nsp13 è uno dei candidati più promettenti per lo sviluppo di terapie pan-coronavirus. [21] I risultati di [22] mostrano che la mutazione P323L presenta una rapida emergenza durante la fase di contenimento e una fase di impennata precoce durante la prima ondata. Queste sostituzioni emergono da varianti genomiche minori che diventano la sequenza dominante del genoma virale. Durante l'infezione, vi è una rapida selezione di L323 nella sequenza dominante del genoma virale ma non di Spike G614. La genetica inversa viene utilizzata per creare due virus (P323 o L323) con lo stesso background genetico.

Figura 13: Overview della mutazioni. (A) Distribuzione con infografica delle 10 proteine con il maggior numero di mutazioni; (B) Distribuzione delle 10 mutazioni più frequenti tra i pazienti; (C) Matrice di cooccorrenza delle 10 mutazioni più frequenti per comprendere quali mutazioni hanno più probabilità di mostrarsi insieme; (D) Top 20 campioni di pazienti che presentano il maggior numero di mutazioni tra la popolazione;



L323 mostra una maggiore abbondanza di RNA e proteine virali e una morfologia della placca più piccola rispetto a P323.

N: La proteina N, il nucleocapside, che costituisce il guscio protettivo dei virus. Secondo una ricerca condotta dalla Pennsylvania State University la sua struttura risulta simile in vari coronavirus e addirittura fra le varianti del Sars-Cov-2.Questo la rende un potenziale bersaglio da colpire attraverso terapie mirate per ridurre infiammazione e sintomi persistenti tipici di chi soffre di long Covid. [23] Seppur la maggior parte degli studi si concentra sulla proteina Spike, in quanto è più esterna, la proteina N funge da rivestimento del genoma del virus ma è racchiusa in uno strato più interno. La sua maggiore profondità fa sì che sia quella in generale più conservata e meno soggetta alle mutazioni che danno luogo alla comparsa di forme virali differenti, le varianti. Dopo che il virus è entrato nell'organismo nel sangue la proteina N, circola liberamente e causa una forte risposta immunitaria con la produzione di anticorpi. [24]

NSP6: La proteina non strutturale 6 è una delle due proteine non strutturali codificate dal gene 11 del rotavirus insieme a NSP5. È una presunta proteina del dominio transmembrana. NSP6 è composto da sei domini transmembrana e una coda C terminale. La proteina NSP6 (ΔSGF), che è sorta indipendentemente nelle varianti Alpha, Beta, Gamma, Eta, Iota e Lambda, si comporta come un mutante con guadagno di funzione con un'attività di chiusura lampo ER più elevata. [25] NSP6 è coinvolta nella formazione delle vescicole a doppia membrana (DMVs), che sono strutture all'interno delle cellule ospiti dove avviene la replicazione del

RNA virale. Queste strutture proteggono il materiale genetico del virus dal sistema immunitario dell'ospite. NSP6, insieme ad altre proteine come NSP3 e NSP4, induce la formazione di queste DMVs, creando un microambiente ideale per la replicazione virale. Diverse varianti del SARS-CoV-2 hanno presentato mutazioni in NSP6. Ad esempio, la delezione Δ106-108 nella proteina NSP6 è una mutazione osservata in varianti come Alpha (B.1.1.7), Beta (B.1.351), e Omicron (B.1.1.529). Queste mutazioni potrebbero alterare la funzione di NSP6, influenzando la patogenicità e la trasmissibilità del virus.

3 Architetture e impostazione del problema

L'analisi dei dati genomici gioca un ruolo cruciale nella comprensione dei meccanismi alla base delle malattie genetiche. Questo campo si occupa dello studio del genoma, l'insieme completo del DNA di un organismo, inclusi tutti i suoi geni. Con l'avanzamento delle tecnologie di sequenziamento del DNA, è diventato possibile generare enormi quantità di dati genomici in tempi relativamente brevi e a costi sempre più ridotti. Questi dati offrono l'opportunità senza precedenti di esplorare la base genetica di malattie complesse, di identificare mutazioni che causano malattie, e di sviluppare trattamenti mirati più efficaci.

I problemi relativi all'utilizzo di dati genomici sono molteplici, di seguito i principali.

Il primo problema è la *vasta quantità di dati generati*. I dati genomici sono estremamente vasti e complessi. Un singolo genoma umano contiene circa 3 miliardi di basi di DNA, e l'analisi di molteplici genomi per studi comparativi o di associazione aumenta esponenzialmente questo volume. Oltretutto, per la genomica, l'acquisizione dei dati è altamente distribuita e coinvolge formati eterogenei. [11]

Il secondo problema è *relativo alla gestione e conservazione dei dati*. Conservare, gestire e accedere efficacemente a grandi set di dati genomici richiede infrastrutture IT dedicate e costose, oltre a specifiche competenze tecniche. [12]

Il terzo problema è strettamente legato *all'analisi e interpretazione dei dati*. L'interpretazione dei dati genomici richiede una comprensione profonda dei meccanismi biologici, oltre alla capacità di distinguere variazioni genetiche rilevanti da quelle che non lo sono. L'identificazione di nuovi marcatori genetici o la comprensione del significato clinico di specifiche varianti rimane una sfida. Per interpretare le sequenze genomiche e rispondere a domande *"Come funziona il DNA? Le mutazioni, cambiamenti di espressione o altre misurazioni molecolari ? "* richiede l'integrazione di competenze nel settore biologico, sistemi di apprendimento automatico su larga scala e informatica infrastruttura in grado di supportare query flessibili e dinamiche per la ricerca di modelli per collezioni di dati con dimensioni molto elevate. [13] [11]

Il quarto problema riguardano tutte le questioni di privacy e di etica. La raccolta, l'uso e la condivisione dei dati genomici sollevano importanti questioni etiche, in particolare riguardo alla privacy, al consenso informato e all'uso dei dati per la ricerca. Assicurare la privacy e la sicurezza dei dati è fondamentale per proteggere le informazioni genetiche degli individui da usi impropri o accessi non autorizzati. [14]

Il quinto problema è strettamente connesso agli *strumenti software.* Lo sviluppo di strumenti software per l'analisi dei dati genomici deve tenere il passo con l'evoluzione delle tecnologie di sequenziamento e con la crescente complessità dei dati. L'efficienza, l'accuratezza e l'usabilità degli strumenti sono critiche. [15]

I problemi appena menzionati si concentrano principalmente nell'ambito informatico/statistico. È importante, però, sottolineare che ci sono diverse conseguenze a livello biologico e medico che non possono essere ignorate. La prima conseguenza è la corrispondenza tra la rilevanza biologica e quella statistica di un gene. La rilevanza biologica di un gene può fornire informazioni preziose per la scoperta di funzioni specifiche del gene, la determinazione di gruppi di geni che contribuiscono alla non sopravvivenza dei pazienti o allo sviluppo di tessuti cancerogeni. Queste informazioni possono influenzare le decisioni cliniche e la cura dei pazienti, rendendo la corretta valutazione della rilevanza biologica un processo fondamentale nell'ambito medico. La seconda conseguenza è derivabile dal problema di inquinamento o contaminazione del campione: questa situazione si verifica quando il campione biologico viene accidentalmente contaminato da batteri, virus o da altri microrganismi durante il processo di prelievo o manipolazione. Questa contaminazione, può portare a risultati di test errati influenzando così la diagnosi e il relativo trattamento del paziente. Per evitare la contaminazione del campione, è importante seguire le corrette procedure di prelievo e manipolazione e utilizzare tecniche di sterilizzazione appropriate.

3.1 Data Problem

L'obiettivo principale dello studio è quello di ottenere una rappresentazione fcgr per ciascun paziente sottoposto al sequenziamento. Le rappresentazioni utilizzate vengono poi passate alle cnn costruita ad hoc per ottenere una corretta classificazione.

I dati delle mutazioni , disponibili sul sito, sono in formato txt. E' stato necessario definire una catena di processi utili alla trasformazioni dei dati per generare le rappresentazioni Frequency Chaos Game Representation (FCGR) per ciascun paziente. Vengono descritti , di seguito, i problemi associati a ciascun passaggio in un'ampia panoramica del flusso di lavoro bioinformatico. E' stato necessario, in primis, trasformare il file *txt* in file *maf e poi in vcf.*

I file *Variant Call Format* (VCF) rappresentano le varianti genetiche individuate attraverso il sequenziamento. Tuttavia, la loro utilità è intrinsecamente legata alla qualità e alla precisione con cui le varianti sono state chiamate. Errori di sequenziamento o limitazioni del software di chiamata delle varianti possono introdurre *inaccuracies* che influenzano tutte le fasi successive dell'analisi.

In seguito, ciascun vcf è stato ordinato rispetto ai cromosomi e successivamente è stato ottenuto l'indice. Durante l'ordinamento e l'indicizzazione, è cruciale mantenere l'allineamento corretto delle varianti con i rispettivi cromosomi e posizioni per evitare errori di interpretazione.

Si procede , poi , alla generazione dei file FASTA attraverso la funzione *FastaAlternateGenerate*, e la successiva estrazione di file FASTA specifici, illustrano la transizione dai dati grezzi di sequenziamento a una forma che può essere utilizzata per la costruzione delle rappresentazioni fcgr. Un file FASTA è un formato di testo semplice e ampiamente utilizzato per memorizzare sequenze nucleotidiche o proteiche. Ogni sequenza in un file FASTA è composta da due parti principali:

- L'intestazione (Header): Inizia con un carattere ">" seguito da un identificatore della sequenza, che può includere informazioni come il nome del gene, l'organismo di origine, e altre annotazioni.

- La sequenza nucleotidica**:** Una stringa di lettere che rappresentano i nucleotidi (A, T, C, G) per il DNA.

Infine si ottiene la rappresentazione scelta, la Frequency Chaos Game Representation (FCGR), che permette di trasformare una sequenza di nucleotidi in un formato bidimensionale che preserva le informazioni sulla frequenza e sulla disposizione dei k-mers (sottosequenze di lunghezza k) all'interno della sequenza originale.

Un k-mer è una sottosequenza di nucleotidi di lunghezza k estratta dalla sequenza originale. Nel contesto del nostro problema, stiamo utilizzando k=7, il che significa che stiamo dividendo la sequenza genomica in tutte le possibili combinazioni di 7 nucleotidi consecutivi. Ad esempio, dato un tratto di sequenza "ATCGTACG", i 7-mer possibili sono "ATCGTAC", "TCGTACG", ecc.

Il numero totale di k-mers possibili per k=7 è dato da 4^k, dove 4 rappresenta il numero di nucleotidi (A, T, C, G). Per k=7, ci sono 16.384 (4^7) possibili k-mers.

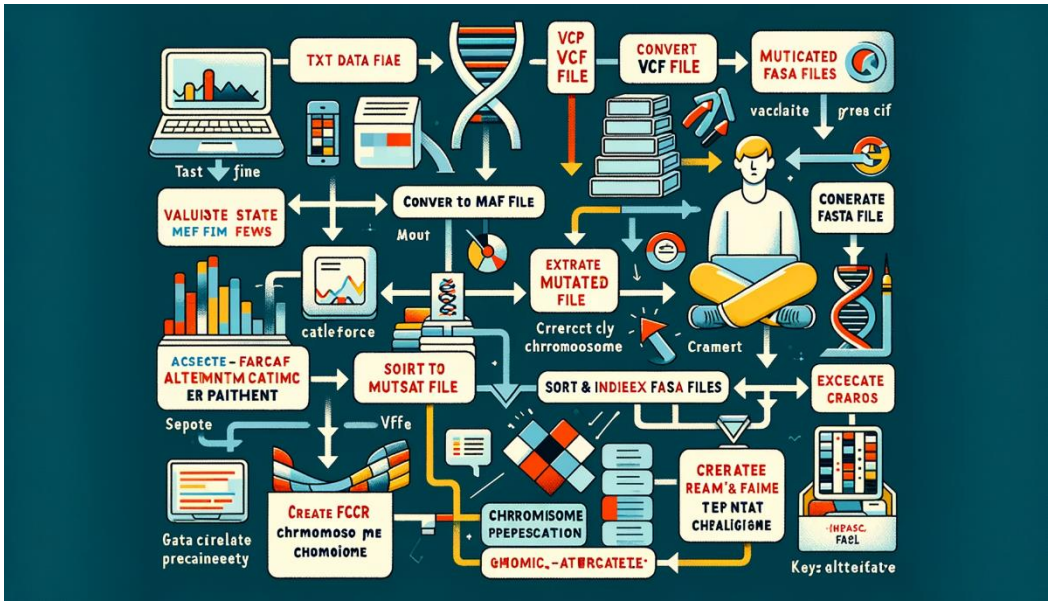


Figura 12: DATA PROBLEM.

3.2 Generazione della Rappresentazione FCGR

La **Frequency Chaos Game Representation (FCGR)** è una tecnica che mappa la frequenza dei k-mers in una matrice bidimensionale. I passaggi per ottenere la FCGR da una sequenza genomica contenuta in un file FASTA sono i seguenti:

1. **Estrazione dei k-mers:**
    - Scorrendo la sequenza genomica nucleotide per nucleotide, vengono estratti tutti i k-mers di lunghezza 7.
    - Per una sequenza di lunghezza N, il numero totale di k-mers estratti sarà N-k+1.

2. **Calcolo della Frequenza dei k-mers:**
    - Una volta estratti i k-mers, si calcola la frequenza di ciascun k-mer nella sequenza. Questo processo produce un vettore di lunghezza 16.384, dove ogni elemento rappresenta il conteggio di un particolare k-mer in quella specifica sequenza.
3. **Mappatura della Frequenza nella FCGR:**
    - La FCGR viene rappresentata come una matrice bidimensionale di dimensione 128x128 (poiché la radice quadrata di 16.384 è 128).
    - Ogni cella della matrice rappresenta la frequenza di uno specifico k-mer, posizionata in base a un'assegnazione deterministica che dipende dall'ordine alfabetico dei nucleotidi nei k-mers. Le posizioni delle celle nella matrice sono determinate dalla sequenza dei nucleotidi, permettendo di preservare le informazioni spaziali all'interno della sequenza.
4. **Normalizzazione della Matrice FCGR:**
    - Per garantire che le rappresentazioni delle diverse sequenze siano comparabili, la matrice di frequenza può essere normalizzata, ad esempio dividendo ogni valore per il numero totale di k-mers nella sequenza, ottenendo così una frequenza relativa.

La rappresentazione FCGR viene ottenuta sia per le osservazioni della classe O che per quelli della classe S.

3.3 Costruzione della CNN

Per la costruzione dell'architettura si utilizza una rete neurale convoluzionale ad hoc.

Prima di tutto si carica un file CSV che contiene i nomi delle immagini e le relative etichette di classificazione (Clade O o Clade S). In base a queste etichette, vengono caricate le immagini dalle directory corrispondenti, ridimensionate a 128x128 pixel e convertite in array numerici. Viene ,quindi, creato un array X contenente tutte le immagini convertite e un array y con le etichette. Queste ultime seguono un ulteriore processo che consente di trasformarle da valori categorici in formato one-hot.

I parametri di input della rete sono : il numero di righe e colonne dell'immagine (128 x 128 pixel) e il numero di classi.

La rete inizia con diversi strati convoluzionali 2D. Ogni strato convoluzionale applica un filtro di convoluzione alle immagini di input per estrarre delle caratteristiche rilevanti:

- ogni conv2D utilizza 4 filtri e un kernel di dimensioni 2 ** level, cioè 8x8 pixel;
- la stride è impostata anch'essa su 2 ** level, ovvero 8, il che significa che il filtro si sposta di 8 pixel alla volta;
- Dopo ogni strato convoluzionale è stata impiegata la funzione di attivazione ReLU (Rectified Linear Unit), scelta per la sua capacità di introdurre non linearità nel modello, migliorando così la capacità di apprendimento del modello rispetto a dati complessi.;

- ciascun strato convoluzionale è seguito da un'operazione di normalizzazione (BatchNormalization), che standardizza l'output dello strato convoluzionale per migliorare la stabilità e la velocità di convergenza durante l'addestramento;
- ogni strato convoluzionale era seguito da un'operazione di **Max Pooling** con finestra **2x2**. Questa tecnica ha ridotto la dimensionalità dei dati, preservando al contempo le informazioni più rilevanti, contribuendo a una maggiore efficienza computazionale e riducendo il rischio di overfitting.
- la rete termina con uno strato Flatten, che converte l'output 2D degli strati precedenti in un vettore 1D, preparandolo per il livello denso finale;
- lo strato finale è un livello Dense con n_classes unità (in questo caso, 2), utilizzando l'attivazione softmax, che converte l'output in una distribuzione di probabilità su ciascuna classe;
- il modello è compilato con la funzione di perdita binary_crossentropy, adatta per problemi di classificazione binaria;
- l'ottimizzatore Adam è stato scelto con un learning rate iniziale di 0.01. Adam è noto per la sua efficienza nella gestione di grandi volumi di dati e per la capacità di adattare dinamicamente il learning rate, contribuendo a una rapida e stabile convergenza durante l'addestramento.
- Il modello è stato addestrato con una batch size di 64, bilanciando l'accuratezza della stima del gradiente con l'efficienza computazionale.

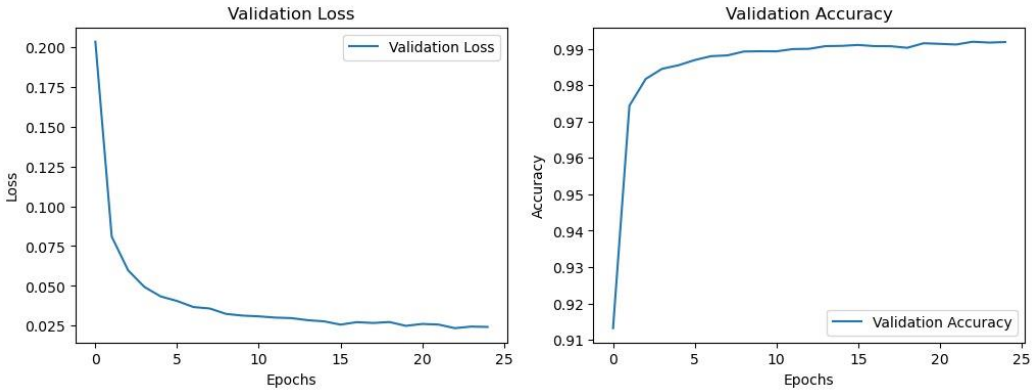
4 Esperimenti e risultati.

Gli esperimenti sono stati condotti su una macchina equipaggiata con un processore Intel Core i9-14900K, che dispone di 24 core suddivisi in 8 Performance-cores (P-core) e 16 Efficiency-cores (E-core), operanti a una frequenza di 6.0 GHz, con supporto per un totale di 32 thread. La macchina è dotata di 32 GB di memoria RAM DDR5 a 7200 MHz e di una scheda grafica NVIDIA GeForce RTX 4070 Ti Super con 16 GB di memoria GDDR6 dedicata. La scheda madre utilizzata è una MSI MAG Z790 TOMAHAWK MAX WIFI, mentre il sistema di raffreddamento è un dissipatore a liquido NZXT Kraken 360. L'alimentazione è garantita da un alimentatore MSI MPG A1000G PCIE5. Il sistema operativo installato è Windows 11 Pro. Per l'esecuzione degli esperimenti è stato utilizzato Python nella versione 3.10.14.

Il numero di istanze è 38050 per la classe O e 15886 per la classe S. L’80% della coorte viene utilizzata per l’addestramento mentre il restante 20% per la fase di testing.

La rete neurale è stata addestrata utilizzando il set di parametri spiegati precedentemente. L'addestramento del modello è stato condotto su 25 epoche. E’ stato scelto per garantire che il modello avesse sufficienti opportunità per apprendere le caratteristiche chiave del dataset senza sovra-adattarsi ai dati di addestramento.

- **Inizio dell'addestramento:** Nelle prime epoche, il modello ha rapidamente migliorato le sue prestazioni, imparando a riconoscere caratteristiche di base. Durante queste fasi iniziali, l'accuratezza sui dati di addestramento è aumentata costantemente.



- **Stabilizzazione:** Intorno alla 15a-20a epoca, il modello ha iniziato a stabilizzarsi, con l'accuratezza che si avvicinava progressivamente al plateau. In questo stadio, il modello ha affinato la sua capacità di distinguere tra le diverse classi, riducendo al minimo la perdita.
- **Fine dell'addestramento:** Alla 25a epoca, il modello ha raggiunto un'accuratezza del 98%. Monitorando la curva della funzione di perdita e dell'accuratezza, è stato osservato che ulteriori epoche non avrebbero comportato significativi miglioramenti, segnalando che il modello aveva raggiunto un buon equilibrio tra bias e varianza.

Il modello CNN ha raggiunto un'accuratezza del 98% nel compito di classificazione, un risultato estremamente positivo che indica un'alta capacità del modello di predire correttamente le classi del dataset.

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

Nonostante la presenza di un dataset fortemente sbilanciato, 70% e 30%, il modello mostra prestazioni eccellenti, con una precisione, recall e F1-score molto elevati per entrambe le classi.

CAPITOLO 4

CONCLUSIONI E SVILUPPI FUTURI

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 (Frequency Chaos Game Representation) tramite immagini. 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. Sono stati testati diversi modelli CNN, dimostrando un'eccellente capacità di generalizzazione del modello, come evidenziato dall'aumento dell'accuratezza durante la fase di validazione. Questo risultato rafforza la fiducia nella capacità del modello di fornire previsioni accurate e affidabili riguardo la struttura tridimensionale delle proteine.

Tuttavia, nonostante i risultati incoraggianti, potrebbero essere necessarie ulteriori ricerche per identificare una rete che possa gestire efficacemente rappresentazioni con un numero maggiore di k-mers.

Gli sviluppi futuri di questo progetto potrebbero esplorare diverse direzioni promettenti:

- L'approccio basato su immagini FCGR potrebbe essere esteso per classificare non solo la proteina Spike, ma anche altre proteine virali. Questo permetterebbe di identificare e distinguere diverse varianti del virus, contribuendo alla comprensione della loro struttura e del loro potenziale impatto sulla patogenicità e la trasmissibilità.

- L'identificazione di modelli CNN ottimizzati per rappresentazioni con un numero maggiore di k-mers potrebbe aumentare la risoluzione e la precisione della classificazione. Questo sviluppo potrebbe migliorare la capacità di distinguere varianti virali con differenze minime nella sequenza proteica.

- L'esplorazione di architetture più avanzate, come reti neurali profonde (deep CNN) o reti residuali (ResNet), potrebbe migliorare ulteriormente le prestazioni del modello, specialmente nella gestione di immagini ad alta dimensionalità derivanti da rappresentazioni FCGR.

- Il modello potrebbe essere utilizzato per monitorare in tempo reale le mutazioni nelle sequenze proteiche, contribuendo alla sorveglianza genomica globale e 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.

RIFERIMENTI E BIBLIOGRAFIA

[1] Mary Fraire, Alfredo Rizzi. *Analisi dei dati per il Data Mining*. Carocci editore, prima edizione, 2011.

[2] Wayne W. Daniel. *Biostatistics: A foundation for Analysis in the Health Sciences.* John Wiley & sons Inc, New York, Sixth Edition.

[3] Chan, E.Y.; Corless, R.M. *Chaos game representation*. SIAM Rev. 2023,65, 261–290.

[4] BARNSLEY MF, VINCE A. *The chaos game on a general iterated function system. Ergodic Theory and Dynamical Systems*. 2011;31(4):1073-1079. doi:10.1017/S0143385710000428

[5] H.Joel Jeffrey. *Chaos game representation of gene structure*, ,Northern Illinois University, DeKalb, IL, USA

[6] Hannah Franziska Löchel, Dominik Heider,

Chaos game representation and its applications in bioinformatics,

Computational and Structural Biotechnology Journal,Volume 19, 2021, Pages 6263-6271, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2021.11.008.

[7] Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M. *Analysis of genomic sequences by Chaos Game Representation. Bioinformatics*. 2001 May;17(5):429-37. doi: 10.1093/bioinformatics/17.5.429. PMID: 11331237.

[8] Gèron A. *Hands on Machine learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools and techniques to build Intelligente Systems.* O' Reilly, Second edition.

[9] Aggarwal, Charu C. *Neural Networks and Deep Learning A Textbook*, 2018 https://www.springer.com/gp/book/9783319944623 (pages 1-167).

[10] Graham MS, Sudre CH, May A, Antonelli M, Murray B, Varsavsky T, et al. The Lancet Public Health. April 12 2021. doi.org/10.1016/S2468-2667(21)00055-4

[11] Stephens, Z.D., et al. (2015). *Big Data: Astronomical or Genomical?.* PLoS Biology, 13(7), e1002195.

[12] Marx, V. (2013). *The big challenges of big data. Nature*, 498(7453), 255-260.

[13] Green, E.D., Guyer, M.S., & National Human Genome Research Institute. (2011). *Charting a course for genomic medicine from base pairs to bedside.* Nature, 470(7333), 204-213.

[14] McGuire, A.L., & Gibbs, R.A. (2006). *Genetics. No longer de-identified*. Science, 312(5772), 370-371.

[15] Fonseca, N.A., Rung, J., Brazma, A., & Marioni, J.C. (2012). *Tools for mapping high-throughput sequencing data. Bioinformatics*, 28(24), 3169-3177.

[16] https://www.sifweb.org/sif-magazine/voci-di-supporto/proteina-spike-di-sars-cov-2

[17] Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR, Mirchandani D, Scharton D, Bilello JP, Ku Z, An Z, Kalveram B, Freiberg AN, Menachery VD, Xie X, Plante KS, Weaver SC, Shi PY. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021 Apr;592(7852):116-121. doi: 10.1038/s41586-020-2895-3. Epub 2020 Oct 26.

Erratum in: Nature. 2021 Jul;595(7865):E1. doi: 10.1038/s41586-021-03657-2. PMID: 33106671; PMCID: PMC8158177.

[18] Frutos R, Yahi N, Gavotte L, Fantini J, Devaux CA. Role of spike compensatory mutations in the interspecies transmission of SARS-CoV-2. One Health. 2022 Dec;15:100429. doi: 10.1016/j.onehlt.2022.100429. Epub 2022 Aug 29. PMID: 36060458; PMCID: PMC9420691.

[19] Lista MJ, Winstone H, Wilson HD, Dyer A, Pickering S, Galao RP, De Lorenzo G, Cowton VM, Furnon W, Suarez N, Orton R, Palmarini M, Patel AH, Snell L, Nebbia G, Swanson C, Neil SJD. The P681H Mutation in the Spike Glycoprotein of the Alpha Variant of SARS-CoV-2 Escapes IFITM Restriction and Is Necessary for Type I Interferon Resistance. J Virol. 2022 Dec 14;96(23):e0125022. doi: 10.1128/jvi.01250-22. Epub 2022 Nov 9. PMID: 36350154; PMCID: PMC9749455.

[20] Newman, J.A., Douangamath, A., Yadzani, S. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nat Commun* **12**, 4848 (2021). https://doi.org/10.1038/s41467-021-25166-6

[21] Yazdi AK, Pakarian P, Perveen S, Hajian T, Santhakumar V, Bolotokova A, Li F, Vedadi M. Kinetic Characterization of SARS-CoV-2 nsp13 ATPase Activity and Discovery of Small-Molecule Inhibitors. ACS Infect Dis. 2022 Aug 12;8(8):1533-1542. doi: 10.1021/acsinfecdis.2c00165. Epub 2022 Jul 13. PMID: 35822715.

[22] Goldswain, H., Dong, X., Penrice-Randal, R. *et al.* The P323L substitution in the SARS-CoV-2 polymerase (NSP12) confers a selective advantage during infection. *Genome Biol* **24**, 47 (2023). https://doi.org/10.1186/s13059-023-02881-5

[23] Kim SM, Kim EH, Casel MAB, Kim YI, Sun R, Kwak MJ, Yoo JS, Yu M, Yu KM, Jang SG, Rollon R, Choi JH, Gil J, Eun K, Kim H, Ensser A, Hwang J, Song MS, Kim MH, Jung JU, Choi YK. SARS-CoV-2 variants with NSP12 P323L/G671S mutations display enhanced virus replication in ferret upper airways and higher transmissibility. Cell Rep. 2023 Sep 26;42(9):113077. doi: 10.1016/j.celrep.2023.113077. Epub 2023 Sep 6. PMID: 37676771.

[24] https://www.wired.it/scienza/biotech/2021/04/15/proteina-n-coronavirus-bersaglio-terapie/

[25] Ricciardi S, Guarino AM, Giaquinto L, Polishchuk EV, Santoro M, Di Tullio G, Wilson C, Panariello F, Soares VC, Dias SSG, Santos JC, Souza TML, Fusco G, Viscardi M, Brandi S, Bozza PT, Polishchuk RS, Venditti R, De Matteis MA. The role of NSP6 in the biogenesis of the SARS-CoV-2 replication organelle. Nature. 2022 Jun;606(7915):761-768. doi: 10.1038/s41586-022-04835-6. Epub 2022 May 12. PMID: 35551511; PMCID: PMC7612910.