UNIVERSITÀ DEGLI STUDI
DI SALERNO

GeneFusion BWT: Feature Extraction e Classificazione su Dati Genomici Reali

Docente / Tutor	**Studenti**
Rocco Zaccagnino	Carmine Calabrese
	Daniele Dello Russo
Alessia Ture	
Gerardo Benevento	

Table of Contents

1 Introduzione e Definizione del Problema

1.1 Il fenomeno della Gene Fusion

Le fusioni geniche sono geni "ibridi" che originano dall'unione di parti di due o più geni originari. Queste anomalie non sono semplici eventi casuali, ma agiscono come veri e propri "motori" (driver) molecolari capaci di innescare la trasformazione maligna e sostenere la progressione tumorale. La loro importanza biologica risiede nell'elevata specificità: poiché le fusioni somatiche si trovano esclusivamente nelle cellule cancerose e non in quelle sane, rappresentano un bersaglio ideale per lo sviluppo di terapie [1].

1.1.1 Significato Clinico e Potenziale Terapeutico

Lo studio di queste sequenze chimeriche riveste un ruolo cruciale nella ricerca oncologica moderna, offrendo applicazioni che vanno dalla diagnosi alla gestione terapeutica:

- **Precisione Diagnostica e Prognostica:** Le fusioni vengono utilizzate come marcatori per confermare sottotipi specifici di cancro. Un esempio storico è l'identificazione del trascritto BCR-ABL1, essenziale per la diagnosi della leucemia mieloide cronica. Alcune fusioni forniscono inoltre indicazioni sulla prognosi, aiutando a prevedere l'aggressività della malattia [1].
- **Monitoraggio della Terapia:** La quantificazione dei livelli di questi trascritti permette di valutare in tempo reale la risposta del paziente e di individuare precocemente l'eventuale perdita di efficacia del trattamento o l'insorgenza di resistenze ai farmaci[1].
- **Target Terapeutico e Repurposing:** La loro unicità permette di progettare farmaci mirati che colpiscono selettivamente la cellula malata, riducendo gli effetti collaterali. Inoltre, la scoperta della medesima fusione in tumori di origini differenti apre la strada all'adozione di farmaci già esistenti per nuove indicazioni terapeutiche[1].

Infine, comprendere i meccanismi di formazione delle fusioni derivanti da riarrangiamenti cromosomici come traslocazioni, delezioni o inversioni, o da errori di trascrizione come il read-through è fondamentale per mappare l'instabilità del genoma tumorale e approfondire la conoscenza della biologia del cancro[1].

1.2 Obiettivi dello Studio

L'obiettivo principale di questo lavoro è creare e testare un sistema di analisi per individuare le fusioni geniche in modo sistematico. Nello specifico, lo studio punta ad applicare la Trasformata di Burrows-Wheeler (BWT) a un dataset controllato, composto sia da sequenze sane che da sequenze chimeriche. Usando la BWT, vogliamo riordinare le informazioni contenute nelle basi del DNA per far emergere pattern strutturali che normalmente non si vedono nelle letture grezze, chiamate reads, prodotte dal sequenziamento. Il cuore della ricerca consiste nell'estrarre da queste trasformazioni dei parametri numerici, o features, capaci di *"fotografare"* le anomalie prodotte dall'unione di geni diversi. Questi dati serviranno poi ad addestrare vari modelli di Machine Learning (ML). L'obiettivo finale è ottenere uno strumento robusto, capace di distinguere autonomamente le vere fusioni che guidano il tumore

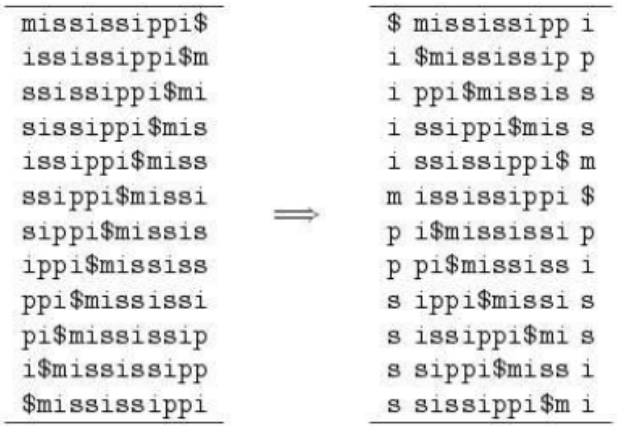

Figure 1: BWT stringa *mississippi*

(driver biologici) dai semplici errori tecnici, che spesso possono portare a false identificazioni durante il processo di sequenziamento.

2 Stato dell'Arte e Background Teorico

2.1 Burrows-Wheeler Transform

La Burrows-Wheeler Transform (BWT) [2] è un algoritmo di trasformazione reversibile dei dati introdotto nel 1994, oggi alla base dei migliori algoritmi di compressione [2]. La BWT non esegue la compressione direttamente, ma agisce come una fase di *pre-processing*: permuta l'ordine dei caratteri di una stringa in modo che simboli con contesti simili si trovino vicini, facilitando la successiva codifica [2].

2.1.1 Calcolo della BWT

L'algoritmo permette la trasformazione di una stringa S di lunghezza `N`, in una trasformata $\hat{S}$ usando i seguenti passaggi [2]:

- **Terminatore:** Alla stringa S viene aggiunto un terminatore `$`.
- **Matrice delle rotazioni:** Si costrisce una matrice $(N+1)$ `x` $(N+1)$ contenente tutti gli shift della stringa S\$.
- **Ordinamento:** Viene effettuato un ordine lessicografico delle stringhe all'interno della matrice.
- **Estrazione:** La stringa $\hat{S}$ sarà l'ultima colonna della matrice.

La figura 1 illustra la matrice ordinata per la stringa *mississippi*, dove la BWT risultante è l'ultima colonna *ipssmpissii*

2.1.2 Reversibilità

La BWT è una trasformazione invertibile: è possibile ricostruire esattamente la stringa originale senza perdita di informazione. Per eseguire l'inversione, è necessario conservare due soli dati [2]:

1. La stringa trasformata $\hat{S}$.
2. La posizione (indice) occupata dal simbolo \$ nella matrice ordinata.

2.2 Varianti della BWT

Per uno studio completo sono state analizzate due varianti specifiche: la Positional BWT e la Reverse-Complement BWT. Queste metodologie permettono di estrarre segnali più precisi, che altrimenti verrebbero "annegati" nel rumore di fondo delle sequenze naturali.

2.2.1 Positional BWT

L'idea alla base della Positional BWT è quella di non considerare la finestra di lettura come un unico blocco uniforme, ma di introdurre una vera e propria risoluzione spaziale. Invece di calcolare un singolo valore di entropia globale, la sequenza viene suddivisa in sottosezioni adiacenti chiamate bin. Calcolando una trasformata indipendente per ogni bin, il modello è in grado di localizzare con precisione dove risieda l'anomalia informativa.
Mentre la BWT classica tende a "diluire" il segnale del disordine, distribuendo l'entropia della giunzione su tutta la lunghezza della finestra, l'approccio posizionale permette di isolarla. In una fusione reale, infatti, ci aspettiamo che l'entropia nei bin centrali, ovvero in corrispondenza del punto di giunzione, sia significativamente più alta rispetto ai bin laterali, che contengono sequenze geniche integre e ordinate.

2.2.2 Reverse-Complement BWT

La Reverse-Complement BWT nasce dalla necessità di assecondare la natura biologica del DNA, che è una struttura a doppio filamento. La trasformata classica analizza tipicamente solo il filamento "Forward", ma in bioinformatica questo approccio può essere limitante. Molti riarrangiamenti strutturali complessi, infatti, lasciano una firma di compressione o dei pattern di ripetizione che risultano molto più evidenti se osservati sul filamento complementare. Applicare la BWT al complementare inverso significa fornire al modello di Machine Learning una "seconda prospettiva" sulla medesima regione genomica. Questo permette di catturare regolarità strutturali che potrebbero apparire mascherate o meno compatte nella lettura standard.

2.3 BWT e Feature Extraction per il Machine Learning

L'utilizzo di sequenze processate tramite BWT in ambito di analisi dati e ML, rispetto all'uso di sequenze grezze, è motivato dalla capacità dell'algoritmo di evidenziare la struttura latente dei dati.

- **Raggruppamento dei contesti:** Nella stringa trasformata $\hat{S}$ i caratteri che condividono lo stesso contesto nella stringa originale sono consecutivi [2].
- **Similitudine Locale:** I simboli tendono a raggrupparsi in gruppi omogenei (Local Similarity) [2].
- **Analisi dei pattern:** La trasformazione rende più evidenti le ripetizioni presenti nella sequenza (analogamente alla Trasformata di Fourier), facilitando l'individuazione di pattern ricorrenti [2].

In sintesi, la BWT riorganizza la sequenza genomica trasformando dipendenze a lungo raggio in adiacenze locali, rendendo i pattern più evidenti e facilmente rilevabili dai modelli di machine learning.

3 Metodologia e Pipeline

3.1 Descrizione dettagliata del dataset

Il dataset di partenza è un dataset basato su dati reali, dove ogni entry rappresenta una sequenza genomica e informazioni utili allo studio di essa. Per ogni entry si ha:

- **sequence:** La stringa di nucleoditi (A, C, G, T).
- **gene1 e gene2:** Identificatori dei geni coinvolti nel fenomeno della fusione.
- **junction_point:** Indice posizionale che indica il punto di fusione.
- **label:** Valore target binario.
  - **Classe Fusion(1):** Sequenza che presenta $gene1 \neq gene2$ e $junction\_point \neq 0$
  - **Classe Non-Fusion(0):** Sequenza che presenta $gene1 = gene2$ e $junction\_point = 0$

3.1.1 Analisi del Bilanciamento

L'analisi preliminare sui dati ha mostrato che le due classi presentano un lieve sbilanciamento (Fig.4):

- **Campioni Non-Fusion**: ≈ 148.000
- **Campioni Fusion**: ≈ 116.000

Nonostante questo divario di circa 30.000 campioni, abbiamo scelto di non alterare artificialmente il dataset (ad esempio eliminando dati reali), ma di gestire lo sbilanciamento integrando due accorgimenti nella fase di training:

- **Stratificazione:** per garantire che i nostri test fossero affidabili, abbiamo diviso i dati tra training e test mantenendo in entrambi le esatte proporzioni del dataset originale. Questo evita che, per puro caso, finiscano troppi pochi esempi di `Fusion` nel set di test.

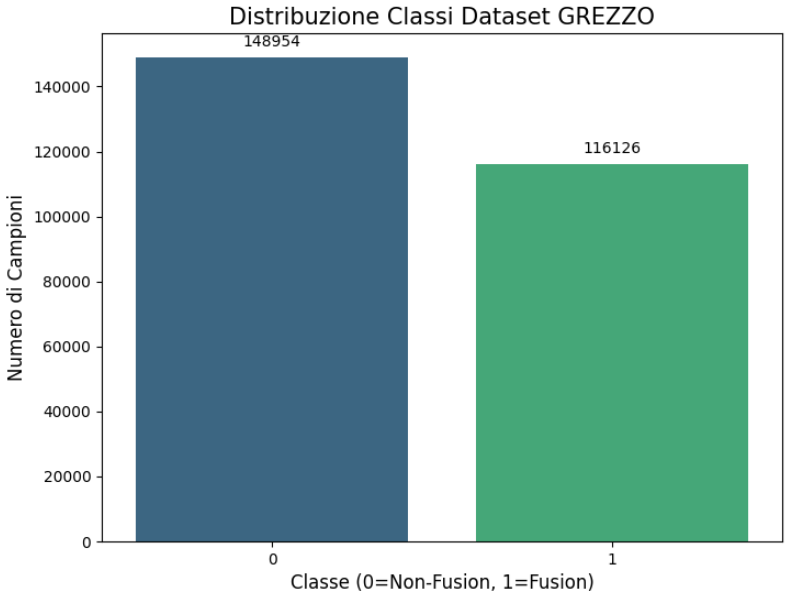

Figure 2: Plot del dataset iniziale

- **Class Weighting:** per impedire al modello di ignorare la classe meno numerosa (e magari ottenere un'accuratezza alta solo indovinando sempre la classe maggioritaria), abbiamo imposto una penalità maggiore agli errori commessi sui campioni `Fusion`. In pratica, sbagliare un caso di fusione "*costa*" di più al modello rispetto a sbagliare un caso normale.

3.2 Preprocessing delle Sequenze

La fase di preprocessing è stata progettata per garantire l'integrità del dato e uniformare gli input per gli algoritmi di Machine Learning.

3.2.1 Rimozione dei Duplicati

Dato l'elevato volume di dati, è stata effettuata una verifica preliminare per rimuovere eventuali ridondanze. Sono state eliminate le righe che presentavano valori identici nella combinazione delle quattro colonne chiave:

- `sequence`
- `gene1`
- `gene2`
- `junction_point`

Questa operazione assicura che il modello non venga valutato su copie esatte di campioni già visti durante il training (vedere Fig.3).

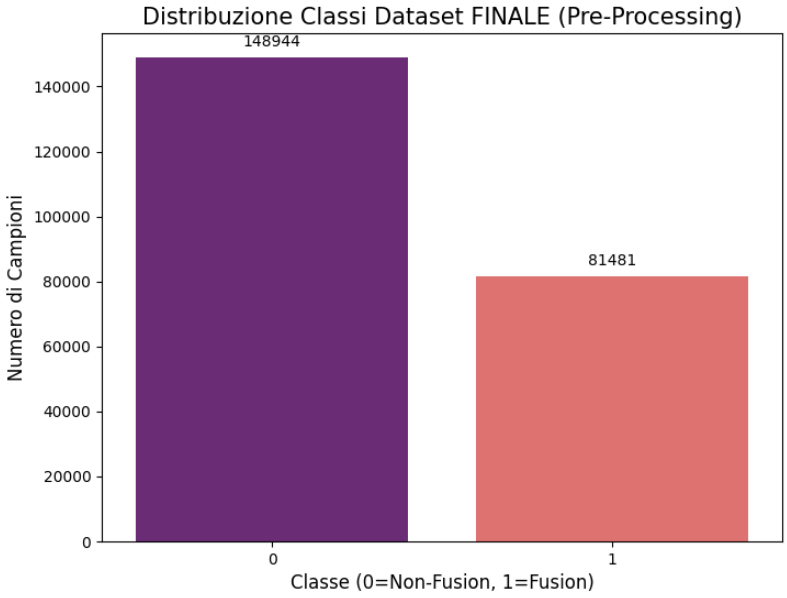

Figure 3: Plot del dataset finale

3.2.2 Segmentazione

Le sequenze genomiche grezze presentano lunghezze variabili che possono portare problemi durante le fasi di training. Per gestire la variabilità è stata prevista una strategia di `windowing` dinamica, che estrae una sequenza di lunghezza pari a **1000 nucleotidi** per ogni entry del dataset. L'algoritmo prevede:

1. **Campioni Fusion (Classe 1):** Per la classe fusion si è scelto di studiare il `junction_point`, ovvero la zona contenente l'informazione discriminante sul processo di fusione.
2. **Campioni Non-Fusion (Classe 0):** In assenza di un punto di fusione si è scelto di usare punti casuali della sequenza per la generazione delle finestre.
3. **Gestione dei Bordi:** L'algoritmo include un controllo sui limiti della stringa, per evitare che venissero presi elementi con posizioni negative o che eccedevano la lunghezza della sequenza.

3.2.3 Data Splitting e Scaling

Una volta ottenute le finestre, il dataset è stato suddiviso in set di addestramento e test con un rapporto 80/20, applicando la stratificazione (come discusso nella sezione 3.1.1). Per il modello `MLP` è stata prevista anche una normalizzazione dei dati per permettere la convergenza del modello.

3.3 Feature Engineering

In questa fase del lavoro, il cuore dell'analisi consiste nel trasformare le sequenze grezze in parametri numerici, definiti features, capaci di addestrare i modelli di Machine Learning. L'approccio adottato è duplice:

- **Analisi diretta:** Alcune caratteristiche vengono estratte direttamente dalla sequenza originale (ACGT) per catturarne la composizione chimica e linguistica.
- **Analisi Strutturale (BWT):** Altre caratteristiche vengono estratte solo dopo aver applicato la Trasformata di Burrows-Wheeler (BWT).

3.3.1 Statistiche di Base

Questo primo gruppo funge da riferimento e si concentra sulla composizione chimica e sull'informazione di base della sequenza. Include il **GC Content**, ovvero la percentuale di basi Guanina e Citosina, e l'**Entropia di Shannon** calcolata sulla sequenza originale. L'obiettivo è verificare se la semplice variazione nella "miscela" di basi sia sufficiente a distinguere una fusione genica, che spesso unisce regioni genomiche con proprietà chimiche differenti.

3.3.2 Caratteristiche Strutturali BWT

In questo sottogruppo, la sequenza viene prima trasformata tramite BWT. Le features estratte includono l'**Entropia della BWT** e l'Analisi dei Run **(Run Length)**, ovvero lo studio della lunghezza delle sequenze di basi identiche che si vengono a creare dopo il riordinamento. Viene inoltre calcolato il Rapporto di Compressione **(Compression Ratio)** tramite l'algoritmo zlib; poiché le fusioni geniche rompono la regolarità naturale del DNA, esse tendono a produrre una BWT più disordinata e meno comprimibile rispetto alle sequenze integre.

3.3.3 Analisi Bio-Linguistica

Questo gruppo tratta il DNA come un linguaggio composto da due "capitoli" che si incontrano nel punto di giunzione. La metrica principale è la **KL Divergence** (Kullback-Leibl), che misura quanto il vocabolario di triplette nucleotidiche cambi drasticamente tra la prima e la seconda metà della finestra analizzata. Viene inoltre monitorata la **densità dei Codoni di Stop**, poiché le fusioni che avvengono fuori registro (out-of-frame) tendono a generare segnali di stop prematuri che non si troverebbero in un gene sano.

3.3.4 Modello Ibrido e Analisi dell'Importanza

L'ultimo gruppo rappresenta l'approccio finale **"Ensemble"**, in cui le caratteristiche più significative dei gruppi precedenti vengono unite in un unico vettore di input. Questo set ibrido combina dati chimici (GC Content), strutturali e bio-linguistici . Attraverso questo modello, è possibile non solo ottenere la massima accuratezza diagnostica, ma anche stilare una classifica dell'importanza delle feature, determinando quali parametri matematici siano i più efficaci nel distinguere i driver biologici reali dagli artefatti tecnici di laboratorio.

3.4 Algoritmi di Classificazione

Nel campo della bioinformatica moderna, gli algoritmi di classificazione sono strumenti essenziali per trasformare milioni di dati grezzi, prodotti dal sequenziamento ad alto rendimento, in informazioni cliniche realmente utilizzabili. Poiché una singola analisi produce milioni di brevi stringhe nucleotidiche, note come reads, diventa indispensabile l'intervento di modelli computazionali capaci di assemblare queste informazioni e identificare sistematicamente le prove di eventuali alterazioni genomiche.

In questo studio, una volta estratte le caratteristiche numeriche (le features) dalle sequenze, il passaggio successivo è stato quello di addestrare dei modelli di Machine Learning affinché imparassero a distinguere autonomamente tra una sequenza sana e una fusione genica. L'obiettivo fondamentale di questi modelli è quello di agire come filtri intelligenti, capaci di riconoscere i reali driver biologici che spingono la cellula verso la trasformazione maligna, isolandoli dai numerosi artefatti tecnici o errori di sequenziamento che possono verificarsi durante i processi di laboratorio.

Per garantire che la valutazione fosse il più possibile robusta e imparziale, non ci siamo affidati a un unico metodo. Abbiamo invece messo a confronto tre diversi modelli di ML **Random Forest, XGBoost e Multilayer Perceptron**, ognuno basato su una logica matematica differente.

3.4.1 Random Forest (RF)

Il Random Forest è stato implementato come modello d'insieme per garantire stabilità contro il rumore di fondo delle *reads*.

- **n_estimators (100):** Il modello utilizza 100 alberi decisionali indipendenti. Questa configurazione permette di ottenere una diagnosi basata sul voto di maggioranza, riducendo l'impatto di singole anomalie causate da errori di sequenziamento.
- **Feature Importance:** La struttura del modello è stata sfruttata per estrarre il peso delle caratteristiche (come la BWT Entropy e la KL Divergence), validando quali parametri matematici siano più indicativi di una fusione reale rispetto a una sequenza integra.

3.4.2 XGBoost (XGB)

L'XGBoost è stato scelto per la sua efficienza nel gestire lo sbilanciamento delle classi, tipico dei dataset oncologici dove le fusioni sono rare.

- **scale_pos_weight:** È l'iperparametro critico impostato dinamicamente in base al rapporto tra campioni sani e fusioni. Questo parametro corregge la tendenza del modello a ignorare i casi positivi, forzando l'attenzione sulle rare sequenze chimeriche.
- **eval_metric (logloss):** Utilizzata per ottimizzare la precisione probabilistica del classificatore durante l'apprendimento di pattern genomici complessi.

3.4.3 Multilayer Perceptron (MLP)

Questa rete neurale è stata progettata per catturare relazioni non lineari tra le *features* estratte dalla BWT e dalla linguistica del DNA.

- **Architettura (64, 32 neuroni):** La rete utilizza due strati nascosti con funzione di attivazione ReLU per processare l'informazione in modo gerarchico.
- **Dropout (0.3):** Impostato al 30% per prevenire l'overfitting. Questa configurazione obbliga la rete a non memorizzare singoli artefatti tecnici di laboratorio, come le chimere generate dalla PCR o dal *template switching* della trascrittasi inversa.
- **Early Stopping:** Una funzione di controllo che interrompe l'addestramento se non si verificano miglioramenti, garantendo che il modello rimanga generalizzabile e non troppo specifico per i dati di training.

4 Risultati Sperimentali

Per uno studio approfondito delle features selezionate, i risultati verranno divisi in gruppi, per valutare le preformance singolarmente e poi nell'interezza del gruppo.

4.1 Metriche di valutazione

Per quantificare l'efficacia dei modelli nella distinzione tra sequenze sane e sequenze chimeriche, sono state utilizzate diverse metriche statistiche derivanti dalla *Confusion Matrix*. L'uso di un set diversificato di metriche è fondamentale data la natura sbilanciata del dataset genomico, dove le classi negative (sequenze sane) prevalgono numericamente sulle positive (fusioni).

- **Accuracy:** La frazione di classificazioni corrette rispetto al totale.
$$\frac{TP+TN}{TP+TN+FP+FN}$$
- **Precision:** Indica l'affidabilità delle predizioni positive, penalizzando i Falsi Positivi (FP).
$$\frac{TP}{TP+FP}$$
- **Recall:** Misura la capacità del modello di individuare tutte le istanze della classe target, penalizzando i Falsi Negativi (FN).
$$\frac{TP}{TP+FN}$$
- **F1-Score:** La media armonica tra Precision e Recall. È la metrica di riferimento per questo studio in quanto bilancia i due aspetti fornendo una valutazione robusta anche con classi sbilanciate.
$$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Confusion Matrix:** Strumento tabellare che confronta le classi predette con quelle reali.

4.2 Esperimento 1: Statistiche di base

Si osservi dalla tabella 1 che il Random Forest ottiene già buoni risultati (F1-Score 0.72), indicando che il **GC Content** è un predittore valido ma non sufficiente. È fondamentale notare il comportamento del MLP (Rete Neurale): con un F1-Score di 0.00 e una Recall di 0.00, il modello non è riuscito a convergere, classificando tutti i campioni come Non-Fusion (classe maggioritaria). Questo evidenzia la difficoltà delle reti neurali semplici nel trovare pattern su poche feature non linearmente separabili.

Table 1: Gruppo 1 (Statistiche di base)

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.7967**	**0.6997**	**0.7445**	**0.7214**
XGBoost	0.5917	0.4508	0.7084	0.5510
MLP (Rete Neurale)	0.6464	0.0000	0.0000	0.0000

4.3 Esperimento 2: Caratteristiche Strutturali BWT

Questo esperimento rappresenta il nucleo della validazione della tesi e l'obiettivo finale del progetto. In questa fase, ci siamo concentrati esclusivamente sulle feature derivanti dalla trasformata di Burrows-Wheeler, escludendo le statistiche biologiche di base, per isolarne il contributo.

I risultati confermano che la BWT è in grado di catturare correttamente le informazioni biologiche latenti delle sequenze: il modello **Random Forest** migliora sensibilmente rispetto alla baseline statistica, raggiungendo un'accuratezza dell'84% e un F1-Score di 0.76. È importante notare anche i progressi degli altri classificatori: le performance di `XGBoost` sono migliorate in termini di Recall e, dato significativo, la rete `MLP` ha raggiunto una convergenza stabile, superando i problemi di apprendimento riscontrati nel primo esperimento.

Di seguito analizziamo nel dettaglio il comportamento di tre varianti della trasformata: Classica, Positional e Reverse-Complement.

4.3.1 Esperimento 2.1: BWT Classica

L'approccio standard si conferma il più efficace in termini di bilanciamento tra precisione e recupero. Con un F1-Score di **0.7672**, questa variante dimostra che il raggruppamento lessicografico dei caratteri è sufficiente a evidenziare i pattern di fusione senza introdurre la complessità computazionale di varianti più elaborate.

Table 2: Risultati Gruppo 2.1 (BWT Classica)

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.8415**	**0.7981**	0.7387	**0.7672**
XGBoost	0.6320	0.4875	**0.7921**	0.6036
MLP (Rete Neurale)	0.6545	0.5224	0.2674	0.3537

4.3.2 Esperimento 2.2: Positional BWT

In questa variante, la trasformata mantiene informazioni sulla posizione originale dei simboli. Nonostante l'aggiunta di informazioni spaziali potesse teoricamente aiutare a localizzare il punto di giunzione, i risultati mostrano una leggera flessione delle performance (F1-Score **0.7465**) rispetto alla BWT classica. Questo suggerisce che, per il compito di classificazione binaria, la presenza o assenza di specifici *b-mers* è più discriminante della loro posizione esatta.

Table 3: Risultati Gruppo 2.2 (Positional BWT)

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.8175**	**0.7334**	0.7601	**0.7465**
XGBoost	0.6167	0.4750	**0.7990**	0.5958
MLP (Rete Neurale)	0.6618	0.5514	0.2342	0.3288

4.3.3 Esperimento 2.3: Reverse-Complement BWT

Considerando la natura a doppio filamento del DNA, è stata testata la BWT applicata anche al filamento complementare inverso. I risultati sono estremamente vicini a quelli della BWT Classica (Accuracy **0.8395**, F1-Score **0.7636**). Questo risultato è molto positivo: indica che l'approccio è robusto e che il segnale della fusione è abbastanza forte da essere rilevato.

Table 4: Risultati Gruppo 2.3 (Reverse-Complement BWT)

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.8395**	**0.7969**	0.7329	**0.7636**
XGBoost	0.6427	0.4966	**0.7757**	0.6056
MLP (Rete Neurale)	0.6677	0.5479	0.3439	0.4226

4.4 Esperimento 3: Analisi Bio-Linguistica

Le feature basate sulla statistica (tabella 5) mostrano performance simili alla baseline. Alcuni appunti fondamentali possono essere:

- Il modello MLP continua a faticare per raggiungere dei risultati accettabili.

- I modelli ad albero sfruttano in maniera più efficace tali features.

Questi appunti permettono di indicare che singolarmente tali features non possono essere indice di individuazione delle sequenze.

Table 5: Gruppo 3 (Analisi Bio-Linguistica).

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.8096**	**0.7120**	**0.7749**	**0.7421**
XGBoost	0.6334	0.4876	0.7210	0.5818
MLP (Rete Neurale)	0.6483	0.5211	0.0683	0.1208

4.5 Esperimento 4: Gruppo Ibrido

La fase finale dell'esperimento ha previsto la selezione delle migliori features degli esperimenti precedenti creando un gruppo di features `ibrido` che utilizzasse differenti proprietà di analisi delle stringhe biologiche per il riconoscimento. Le features scelte dei vari gruppi sono:

- **Gruppo 1:** È stata selezionata solo la `GC Content`, l'entropia è stata scartata per dare spazio all'entropia calcolata sulle stringhe trasformate dalla BWT.
- **Gruppo 2:** Essendo la BWT argomento centrale del lavoro, si è preferito non scartare features di tale gruppo. In seguito all'analisi comparativa, è stata selezionata la **variante Classica** come architettura di riferimento, in quanto ha dimostrato il miglior trade-off tra performance dei modelli e complessità dell'algoritmo.
- **Gruppo 3:** Per il terzo gruppo sono state usate tutte le features, per studiarne il comportamento in simbiosi con le altre selezionate.

L'esperimento ha prodotto buoni risultati, proclamando il `Random Forest` come miglior modello per classificare le sequenza. Altra nota fondamentale, l'uso di sequenza ibride ha portato miglioramenti nelle performance di tutti i modelli scelti (vedere tabella 6).

Table 6: Gruppo 4 (Gruppo Ibrido).

Modello	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	**0.8904**	**0.8969**	**0.7796**	**0.8341**
XGBoost	0.7124	0.5683	0.7776	0.6566
MLP (Rete Neurale)	0.7193	0.6417	0.4670	0.5406

4.6 Analisi dell'importanza

Un'ulteriore analisi sviluppata è stata fatta sull'importanza delle features, un valore calcolato dal `Random Forest` che permette di inferire quali features hanno inciso di più sul riconoscimento finale delle sequenze.

Dall'esecuzione finale è stata fatta un'ulteriore analisi sulle features calcolate, tale analisi è stata condotta tramite il valore dell'importanza delle features calcolata dal modello `Random Forest`. Tale analisi ha fatto emergere che:

- La **KL Divergence** ha un impatto molto alto se usata in combinazione con altre features.
- La trasformata **BWT** è stata cruciale nel processo di riconoscimento delle sequenza.
- L'uso di features eterogenee tra di loro permette ai modelli di imparare caratteristiche differenti delle sequenze permettendo un apprendimento migliore delle caratteristiche.
- L'uso di modelli complessi non garantisce una corretta convergenza e riconoscimento delle stringhe, ma il lavoro fondamentale è la ricerca delle features.

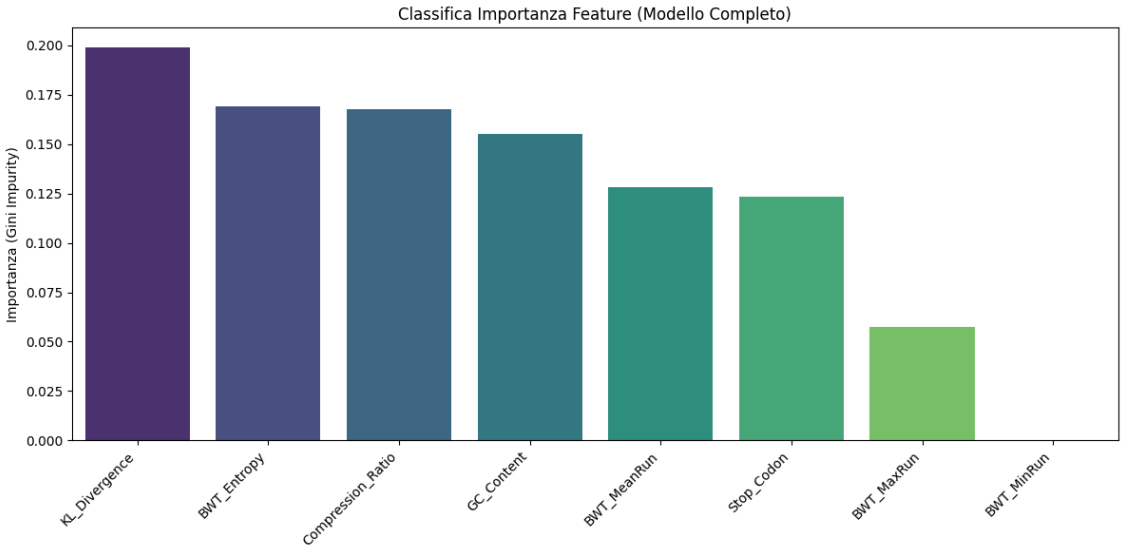

Figure 4: Plot dell'importanza delle features

5 Conclusioni e Sviluppi Futuri

5.1 Sintesi del Lavoro e Validazione dell'Ipotesi

Il presente studio aveva l'obiettivo di validare l'efficacia della **Burrows-Wheeler Transform (BWT)** per l'estrazione di features per classificare eventi di Gene Fusion utilizzando algoritmi di Machine Learning, cercando soluzioni alternative agli approcci classici basati sull'allineamento delle sequenza, cercando di velocizzare il processo. I risultati sperimentali hanno validato l'ipotesi:

- **Superiorità delle Feature Strutturali:** Valutando singolarmente i gruppi di features, escluso l'esperimento finale, le features estratte dalle trasformate BWT hanno prodotto risultati migliori.
- **Successo del Modello Ibrido:** La combinazione di feature eterogenee ha permesso al classificatore **Random Forest** di raggiungere un'accuratezza del **89.04%** e una precisione del **89.69%**. L'analisi della *Feature Importance* ha confermato che variabili algoritmiche come la *Kullback-Leibler Divergence* e la *BWT Entropy* sono predittori più forti rispetto ai marcatori biologici tradizionali.

5.2 Limitazioni e Analisi Critica

Nonostante i risultati promettenti, l'analisi ha evidenziato alcune criticità:

- **Fallimento delle Reti Neurali Semplici (MLP):** I modelli Multi-Layer Perceptron hanno mostrato difficoltà di convergenza, specialmente sui gruppi di feature singoli, collassando spesso sulla classe maggioritaria (Recall ≈ 0). Questo suggerisce che, per dati tabellari strutturati estratti da sequenze, gli algoritmi basati su alberi decisionali (Random Forest, XGBoost) rimangono la scelta più robusta ed efficiente rispetto alle reti fully-connected di base.
- **Trade-off Precision/Recall in XGBoost:** L'utilizzo aggressivo del *Class Weighting* per gestire lo sbilanciamento dei dati ha portato XGBoost a massimizzare la Recall a discapito della Precisione, generando un numero di falsi positivi superiore rispetto al Random Forest. In un contesto clinico reale tale comportamento avrebbe bisogno di un ulteriore controllo umano.

5.3 Sviluppi Futuri

Alla luce dei risultati sperimentali, si identificano alcuni sviluppi futuri per il progetto:

1. **Utilizzo di varianti della BWT:** La variante classica della BWT ha prodotto ottimi risultati, si potrebbe pensare di implementare trasformate BWT differenti che permettono di estrarre informazioni differenti dalle sequenze. Ad esempio si potrebbe pensare di effettuare uno studio approfondito sulla `Rerverse-Complement BWT` essendo una tipologia che ha data buoni risultai, molto vicini alla tipologia classica.
2. **Validazione su dati specifici:** Il modello ibrido è stato addestrato su un dataset reale generico, si potrebbe pensare di effettuare tali esperimenti su dataset patologici specifici per valutarne la performance finale.

Bibliografia

[1] MJ Annala et al. “Fusion genes and their discovery using high throughput sequencing.” In: *Cancer letters* 340.2 (2013), pp. 192–200.

[2] Elad Verbin Shir Landau. “The Burrows-Wheeler compression algorithm is even better than what you have thought.” In: (Apr. 2005).