Statistica e Analisi dei dati: Quantheimer, un'analisi statistica e quantistica sull' Alzheimer

Vincenzo Danese
Dipartimento informatica, Università degli studi di Salerno

Giovanni Russo
Dipartimento informatica, Università degli studi di Salerno

(Prof. Luigi Di Biasi,
Prof. Stefano Cirillo)

(Dated: January 28, 2026)

Questo studio affronta la diagnosi preclinica della malattia di Alzheimer integrando analisi statistica multivariata e Quantum Machine Learning (QML). Analizzando un dataset eterogeneo di 192 individui di mezza età, comprendente profili genetici, psicometrici e biomarcatori ematici, la ricerca mira a identificare fenotipi di rischio latenti.

L'approccio classico, mediante clustering K-Means, ha isolato configurazioni cliniche in cui depressione e coping evitante anticipano il declino cognitivo, talvolta mascherato dalla riserva cognitiva. Parallelamente, l'indagine quantistica ha confrontato Fidelity Quantum Kernels (FQK) e Projected Quantum Kernels (PQK), dimostrando la maggiore robustezza del PQK nel mitigare la concentrazione esponenziale dei dati.

L'esecuzione sperimentale su hardware reale IBM ha generato raggruppamenti topologicamente coerenti con le simulazioni ma strutturalmente divergenti dai cluster classici, come evidenziato da bassi indici di sovrapposizione. In conclusione, il calcolo quantistico emerge come strumento complementare promettente, capace di rivelare correlazioni non lineari tra riserva cognitiva e patologia che sfuggono alle metodologie tradizionali

Repository: Link GitHub alla repository

1. INTRODUZIONE

La malattia di Alzheimer è una patologia neurodegenerativa riconosciuta dall'Organizzazione Mondiale della Sanità come una priorità globale di salute pubblica. Essa rappresenta la causa più comune di demenza, essendo responsabile di circa il 50%–75% dei casi totali. Il quadro clinico si manifesta tipicamente come un declino cognitivo progressivo che, partendo spesso da deficit della memoria episodica, arriva a compromettere l'indipendenza e le attività della vita quotidiana del paziente. Dal punto di vista fisiopatologico, la malattia è caratterizzata dall'accumulo di placche di peptide β-amiloide ($A\beta$) e grovigli neurofibrillari di proteina tau iperfosforilata. Attualmente, la ricerca si sta focalizzando sull'identificazione di una fase preclinica prolungata [3].

L'obiettivo principale del presente studio è identificare, attraverso l'analisi approfondita di un dataset genetico quali siano le caratteristiche comuni e i fattori predittivi associati a una maggiore suscettibilità alla malattia.

Nello specifico, l'indagine è stata guidata da tre domande di ricerca (*Research Questions*, RQ), strutturate per esplorare sia metodologie classiche che approcci innovativi basati sul calcolo quantistico:

RQ1: Esiste una correlazione significativa tra una diagnosi di depressione e una diagnosi positiva di demenza?

RQ2: È possibile identificare pattern alternativi o non convenzionali utili ai fini della diagnosi precoce?

RQ3: Quali differenze emergono dal confronto tra i risultati ottenuti mediante tecniche di clustering tradizionale e quelli derivanti da algoritmi di clustering quantistico?

2. STRUTTURA DEL DATASET

Il dataset analizzato è costituito da una coorte di **192 individui sani**, appartenenti alla fascia della "mezza età" (tra i 50 e i 63 anni). Si tratta di soggetti che, pur non presentando ancora sintomi di demenza, si trovano nell'età critica in cui i processi di invecchiamento cerebrale e i fattori di rischio iniziano a interagire in modo significativo [2].

La struttura dei dati si articola su due livelli di approfondimento:

1. Il profilo base (disponibile per tutti i 192 partecipanti)

Per l'intero campione è stata raccolta una vasta gamma di informazioni che copre tre aree principali:

- **Genetica:** Sono stati isolati dati relativi a due geni specifici, *APOE* [4] (è una glicoproteina la cui funzione primaria è trasportare lipidi nel cervello e nel sistema periferico) e *PICALM* [1] (una proteina adattatrice fondamentale per l'endocitosi mediata da clatrina. In termini semplici, regola il "traffico" in entrata e uscita dalle cellule: aiuta a formare

le vescicole che trasportano molecole dalla superficie cellulare al suo interno). Questi marcatori sono noti in letteratura per aumentare la suscettibilità alla malattia di Alzheimer a esordio tardivo.

- **Psicometria:** Ogni partecipante è stato sottoposto a una batteria di test per valutarne le capacità cognitive e il profilo psicologico, includendo personalità, tono dell'umore e le strategie individuali di gestione dello stress, quali:

  RPM: test d'intelligenza fluida, misura il grado di ragionamento logico di un individuo, valori alti indicano migliori capacità.

  BDI: risultati del test di Beck per la depressione, più è alto il valore più è grave la malattia.

  NEO: rappresenta le 5 aree della personalità.

  MINI-COPE: un questionario psicologico per valutare come le persone rispondono a situazioni di stress, valutato in 14 stati a gruppi di due domande per valutare 3 macro aree.

  CVLT: è uno dei test neuropsicologici più famosi e utilizzati al mondo per valutare la memoria verbale episodica e la capacità di apprendimento.

- **Psicometria meta dati:** A completare il quadro vi sono le informazioni riguardanti i dati psicosociali quali:

  SES: Indicatore dello stato socio-economico basato su scale standard (come l'indice di Hollingshead).

  Eucation: rappresenta il grado di istruzione 1 sono le superiori; 2 post diploma o uni incompleta; 3 laurea.

  AUDIT: stato alcolemico di una persona, più è alto e più indica una dipendenza da alcool.

2. Il profilo avanzato (disponibile per un sottogruppo di circa 79 partecipanti)

Per una parte selezionata della coorte, l'indagine ha integrato due ulteriori tipologie di dati biologici e funzionali:

- **Biomarcatori ematici:** Sono stati effettuati esami del sangue per analizzare l'emocromo completo, il profilo lipidico e l'eventuale presenza del virus HSV (herpes).

3. METODOLOGIA

In questo capitolo presenteremo le varie metodologie di ricerca sul dataset.

3.1. Preprocessing

A seguito della visualizzazione del dataset sono state effettuate le seguenti operazioni:

Calcolo NA: sono stati identificati tutti i valori NA (Not Avaliable).

Calcolo Metriche Statistiche: quali media, deviazione standard ecc.

Encoding dei dati: le variabili testuali sono state mutate in variabili numeriche[1].

Data Imputation: là dove ritenuto necessario sono stati generati i valore a sostituzione degli NA.

Rimozione: sono state cancellate righe o colonne ritenute non necessarie ai fini dell'analisi.

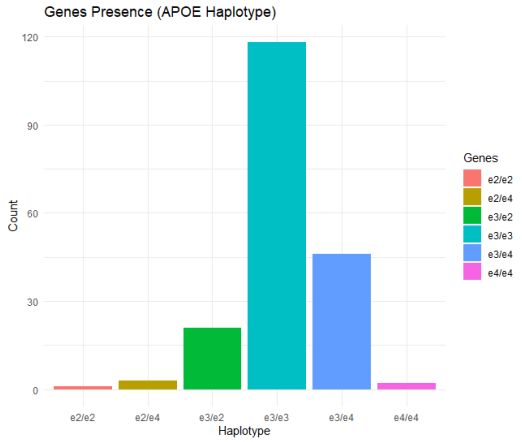

FIG. 1: Istogramma mostrante la distribuzione dei geni APOE.

3.2. Esplorative Data Analysis EDA

Successivamente è stata effettuata una analisi esplorativa dei dati che ha dato origine ai seguenti grafici:

Box plot: al fine di osservare come sono distribuiti i dati in modo compatto.

Density plot: per verificare le simmetrie dei dati e per il confronto diretto di gruppi di dati.

Quantile: per comprenmdere dpve si posizionano i singoli dati.

Correlation heatmap: per visualizzare immediatamente la correlazione tra coppie di dati.

---

[1] Sempre tramite encoding è stata ridotta la dimesionalità di MINI-COPE

Bivariate Analysis: per osservare l'intrazione tra due variabili. [2]

3.3. Riduzione della Dimensionalità

Durante questa fase è stata ridotta la dimensionalità del dataset attraverso la **Principal Component Analysis (PCA)**, in particolare della **CVLT** e dei biomarcatori ematici. Successivamente è stata effettuata una comparazione delle distribuzioni dei dati del dataset con i dati dei 75 pazienti che non hanno acconsentito alla divulgazione dei dati inerenti alle analisi del sangue.

3.4. Fase Finale

A seguito della riduzione della dimensionalità, sono stati generati generati dei *correlation plot* per verificare se ci fossero ulteriori correlazioni e se fosse possibile eliminare altre colonne. Successivamente con il metodo *elbow* e della *silhouette* sono stati definiti il numero di cluster. Infine, sono state applicate misure di qualità dei cluster per verificarne l'ottimalità. Una volta verificato che i cluster fossero ottimi abbiamo Risposto alla **RQ1** e alla **RQ2**.

3.5. Quantum

Al termine dell'analisi classica è stata effettuata l'analisi del clustering quantistica secondo i metodi descritti in *Quantum kernel methods under scrutiny: a benchmarking study* [5]. Per rispondere alla **RQ3**.

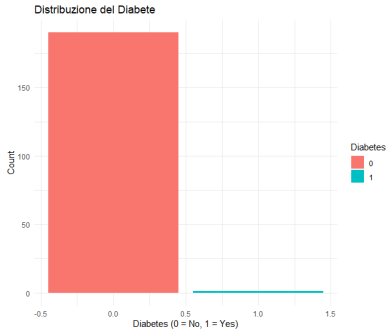

FIG. 2: Distribuzione classe diabetici

4. PREPROCESSING E ANALISI DEI DATI

Per poter utilizzare il dataset per le nostre analisi, abbiamo dovuto effettuare una fase di preprocessing.

[2] Necessaria a causa delle forti correlazioni osservate dagli heatmap di correlazione

Il dataset si compone di quasi tutte le variabili di tipo character, anche per valori numerici, complicando la struttura delle analisi successive. Pertanto, sono stati convertiti i dati necessari in valori numerici. Il passo successivo è stato quello di generare un dataframe con i valori delle metriche statistiche per ogni feature, con alcune di esse aventi **deviazione standard** troppo bassa, che saranno discusse più avanti.
Escludendo i valori delle analisi del sangue che verranno trattati successivamente, alcuni campioni presentavano valori NA; quindi, abbiamo adottato due strategie:

1. **Imputazione tramite mediana:**
   - CVLT class
   - Smoking status
   - hypertension
   - other disease
   - Mini-cope class
   - thyroid disease
   - learning deficits
2. **Imputazione tramite min**
   - education
   - allergies
   - ibuprofen intake

Per finire la fase di imputazione, un paio di campioni sono stati eliminati poiché non erano presenti né i valori psicologici né i valori del sangue.

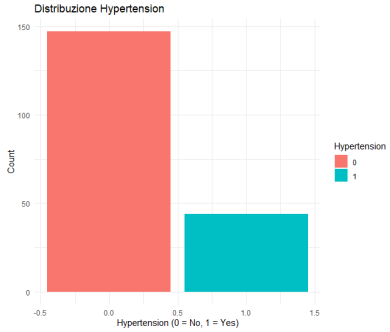

FIG. 3: Classe malati di ipertensione

4.1. Encoding dei dati

Il dataset presenta alcuni dati in formato stringa, quali **APOE** e **PICALM** che si riferiscono ai geni espressi negli individui, per questo motivo si è proceduto nel seguente modo:

- **APOE.** La feature riguardante l'**APOE** si divide in tre elementi. **APOE-rs_xxxxxxx** che indica la

coppia di nucleotidi presente nei locus alla posizione *xxxxxxx* e la codifica vera e propria dei nucleotidi che rappresentano l'allele[1] della proteina espressa. Tutte le features riguardanti il gene **APOE** sono stati mappati secondo una scala numerica, da 1 a 4.

- **PICALM.** Per questo gene si è deciso di usare il one-hot encoding con n-1 gradi di libertà, rappresentando implicitamente quella mancante.

La fase di encoding si è conclusa con la ridefinizione del test **Mini-cope**, qui riportato, su 14 valori che rappresentano il raggruppamento a 2 a 2 del questionario del test. Per tanto, abbiamo sommato e calcolato la media, dividendo il test su 3 gruppi,in riferimento alla struttura del test psicologico. Le colonne Mini-cope 1-14 sono state ridotte in tre gruppi più compatti:

Cope Active:: il test riguarda la sfera cognitiva dell'agire, pianificare e cercare consigli pratici al fine di risolvere un problema da parte del paziente.

Cope Emotive: è legato a quella che è la gestione della sfera emotiva di una persona, come agisce sotto stress o affrontare problema.

Cope Avoidant: si lega all'evitare il problema, bere, negare, incolparsi. Questo gruppo correla spesso con declino cognitivo e depressione.

Cosi facendo abbiamo codificato meglio l'informazione del test raggruppandone i risultati e riducendo la dimensionalità di un fattore lineare di 11.

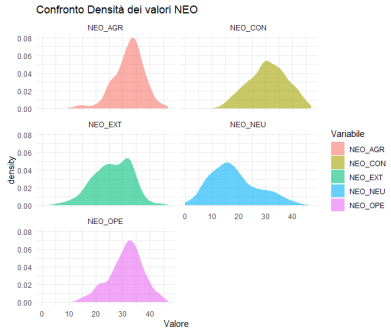

FIG. 4: Grafico che mostra le distribuzione dello spettro della personalità.

5. EXPLORATIVE DATA ANALYSIS

Nella fase esplorativa ci si è concentrati maggiormente sulle distribuzioni delle features presenti, mediante grafici

[1] Ciascuno dei due o più stati alternativi di un gene che occupano la stessa posizione (locus) su cromosomi omologhi e che controllano variazioni dello stesso carattere.

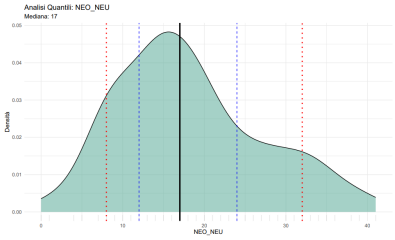

FIG. 5: Distribuzione NEO NEU.

di densità, quantili, boxplot e qualche istogrammi.
Durante questa fase sia partiti con verificare le distribuzione dei geni dell'**APOE**. La distribuzione dei genere segue tutto sommato una . Il valore preponderante risulta essere la coppia $\epsilon 3$ come si evince dalla figura 1 mostrata, con il secondo valore preponderante che presenta nella coppia un $\epsilon 4$ che risulta essere un fattore di rischio per l'alzheimer.

In questa fase si sono confermate molte delle osservazioni fatte durante la fase di studio delle metriche statistiche, riportiamo alcuni esempi.
Diabete e **Ipertensione** risultano essere sbilanciate, come riportato in figura 2 3. In modo analogo anche le features: **education, learning_deficit, other_disease, thyroid_disease, smoking status e allergies** hanno una bassa deviazione standard e valori molto sbilanciati verso un valore o addirittura quasi o completamente costanti come è risultato per il diabete.

Per tanto, dopo un'approfondita analisi sull'importanza di queste features in comparazione con il grafico di correlazione, si è deciso di eliminare le features citate. La ragione sottostante risiede nella necessità di ridurre la dimensionalità del dataset, siccome esso contiene all'incirca 85 colonne, un numero troppo che ci farebbe ricadere nella curse of dimensionality. Si è proseguito con l'analizzare le colonne riguardanti le cinque aree della personalità, indicate dal prefisso **NEO**(Fig 4).
Ciò che ne è risultato è riportato di seguito:

- **NEO_EXT** La distribuzione mostra quasi un plateau intorno la mediana, mostrando valori stabile e persone non troppe sbilanciate su questo ver-

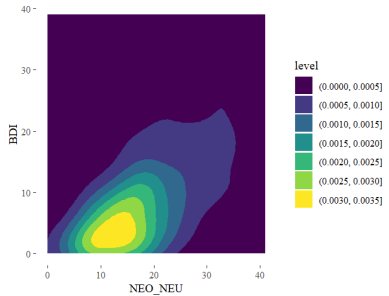

FIG. 6: KDE relativo a NEO_NEU e BDI

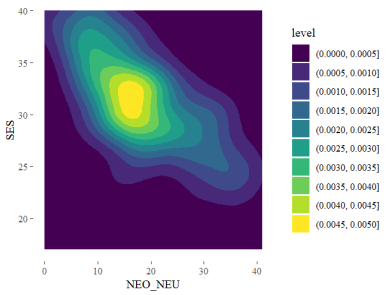

FIG. 7: KDE relativo a NEO_NEU e SES

sante.

- **NEO_AGR** Mostra un piccolo elevato al centro con poca varianza e valori molto elevati, suggerendo che tutti i pazienti sono gentili ed empatici.
- **NEO_CON e NEO_OPE** Entrambe le variabili indipendenti seguono una distribuzione molto simile ad una normale.
- **NEO_NEU** Si nota come una sorta di normale ma con una coda che si allunga molto verso destra, suggerendo molte persone siano molto nevrotiche. L'analisi successiva effettuata con i quantili (Fig 5), ci dice che il 25% risulta essere molto nevrotico.

5.1. Analisi Bivariata

Per approfondire le evidenze emerse dalla heatmap, è stata condotta un'analisi della densità Kernel sulle relazioni più significative.

Il grafico in figura 6 mostra una chiara distribuzione ellittica orientata positivamente: l'addensamento dei punteggi lungo la diagonale conferma una forte correlazione positiva, indicando che all'incrementare dei tratti di nevroticismo corrisponde un aumento coerente della sintomatologia depressiva.

Al contrario, la relazione tra NEO_NEU e SES 7 evidenzia un pattern speculare con pendenza negativa. La concentrazione della densità nel quadrante superiore-sinistro e il suo sviluppo verso il basso-destra attestano una solida correlazione negativa.

6. RIDUZIONE DIMENSIONALITÀ

Una volta conclusa quella che è la fase preliminare di esplorazione dei dati, è stato necessario ridurre ulteriormente la dimensionalità del dataset, arrivando in questa fase a ~ 70 colonne, ancora troppe per poter applicare le tecniche di machine learning.

Come accennato nella sezione riguardante la struttura del dataset, il dataset è composto da un campione di 192 pazienti ma di questi solo 79 aventi le analisi del sangue. Le colonne dei test sanguigni sono ~ 30 il che, visto il campione ridotto, 79 su 192, ci apre la strada a due possibilità in un analisi di getto. La prima è imputare una matrice 113×30 di dati imputati, il che potrebbe sovrastare la bontà dei dati reali. La seconda sarebbe eliminare le analisi del sangue visto il campione ridotto e guadagnare una trentina di colonne. Ma prima di rispondere a questo dilemma, si è cercato di ridurre lo spazio delle features sul test CVLT e quelle delle analisi del sangue per decidere in seguito come procedere.

Ciò che abbiamo ricavato dalla heatmap di correlazione delle variabili 8 vi è una forte correlazione fra quasi tutte le componenti che compongono il test **CVLT**, per questo motivo, abbiamo applicato una PCA su tutte le componenti della CVLT tranne la componente 9 che abbiamo tenuto fuori per verificare quante parole i pazienti inventano durante il test per preserva l'informazione.

Invece la componente 12 e 13 sono state eliminate poiché risultano essere costanti. La PCA sulle componenti della CVLT 9 con solo due componenti ci permette di catturare il 73% dell'informazione originale, andandoci a fermare sul gomito più stretto ed ottimizzando lo spazio delle features, abbiamo cosi ridotto la cardinalità delle variabili indipendenti di un ulteriore dieci taselli.

Terminata la Principal Component Analysis sulle colonne della **CVLT**, è stato il turno delle analisi del sangue. L'analisi condotta sulla varianza riguardante i valori completi dell'emocromo 10, mostra quasi una piatta, per ridurre cosi tante variabili in uno spazio sufficiente da catturare almeno il 60% dell'informazione è richiesto un minimo di cinque dimensioni, rispetto al CVLT che ne sono bastate due siccome la prima componente superava già di suo la soglia minima di informazione necessaria per preservarla. Visto che il nostro obiettivo in questa fase era ridurre il più possibile la dimensionalità del dataset, abbiamo scelto il numero minimo di dimensioni necessarie al fine di non perdere informazione.

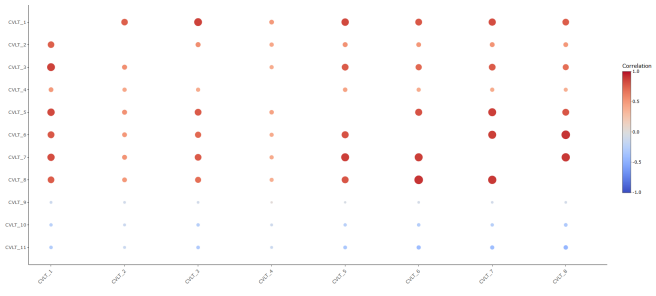

FIG. 8: Grafico della correlazione ristretto sulle componenti CVLT

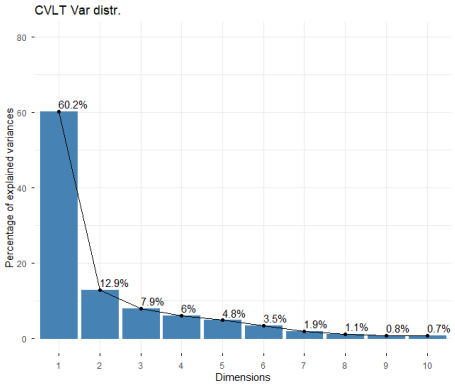

FIG. 9: Distr. varianza delle componenti di CVLT

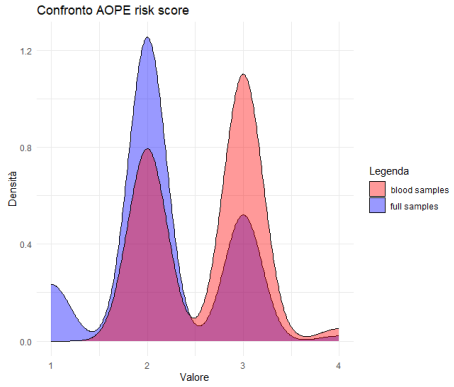

FIG. 11: Confronto colonna APOE risk campione std con campione blood test

6.1. Scelte finali sui blood test

L'approccio adottato dopo aver applicato la PCA anche sui test sanguigni, come riportato nelle sezioni precedenti, è stato il seguente; Per cercare di mantenere queste features importanti ma allo stesso tempo non distorcere i dati presenti, è stato applicato un confronto tra i due dataset (192 samples e 79 samples), tramite metriche statistiche e grafici delle distribuzioni sovrapposti fra loro. Il sotto campione risulta mantenere la stessa struttura e andamento del campione originale analizzato, tranne per alcune coppie analizzate. Discutiamo qui solamente le scelte fatte per queste colonne, siccome si tratta di studiare 18 grafici e non posso essere riportati tutti di seguito, per i grafici non analizzati in questa sessione, si rimanda al link Github della repository.

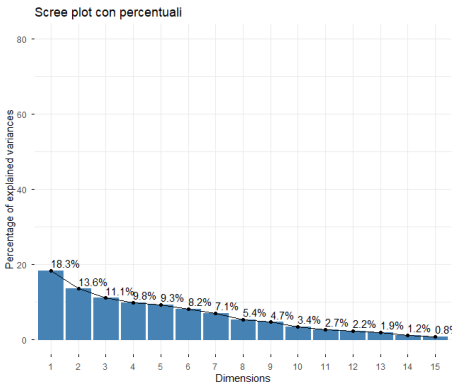

FIG. 10: Distr. varianza delle componenti del blood test

APOE risk: Come riportato in figura 11, la distribuzione del sub-sample rispetto al sample completo, mostra un marcato aumento dei valori di rischio 3, con una riduzione del fattore 2 ma non con una riduzione drastica a nostro avviso, visto il maggior interesse nel preservare i valori di alto rischio genetico per il nostro studio. Per tanto anche se le distribuzione non sono preservate, preferiamo mantenere questa colonna, proseguendo con l'idea di lavorare solo con i 76 campioni.

NEO NEU: L'ultimo caso riguarda quello mostrato in figura 12, il campione base mostra un andamento quasi come una chi-square, con una forte coda verso destra, indicando molti pazienti che soffrono di nevrotismo. Il sotto campione analizzato perde buona parte di questa coda andando ad attenuarsi nel suo centro, ma seguendo tutto sommato quella del campione originale, alcune valutazione tramite la heatmap di correlazione, mostra come questa variabile sia molto correlata positivamente con **BDI(back depression inventory)** e una forte correlazione negativa con **SES**, per questo motivo essa non pregiudica la nostra idea. Ma vista la correlazione citata è stata deciso di escludere questa feature dal dataset, venendo preservata indirettamente nelle due precedenti.

La scelta finale, è stata quella di lavorare solo sul sotto campione che si è ridotto nel corso del processo a 76 campioni, cosi da raggiungere anche una riduzione del campione che è necessaria al fine di essere testato su un computer quantistico (simulato/reale).
Il termine di questa analisi ci ha condotto ad una riduzione significativa dello spazio delle variabili indipendenti, raggiungendo il valore di 21 variabili da usare nel clustering, tagliando fuori alcune colonne, che sono state utilizzate per analizzare i gruppi generati dall'algoritmo di clustering, data la loro struttura che potrebbe inficiare i cluster. Le colonne usate solo per le post analisi sono le seguenti:

- **APOE risk**
- **Dementia history parents**
- **Sex**
- **PICALM_rs3851179_G/A**

Conclusa questa analisi, si è passati alla creazione dei gruppi di cluster.

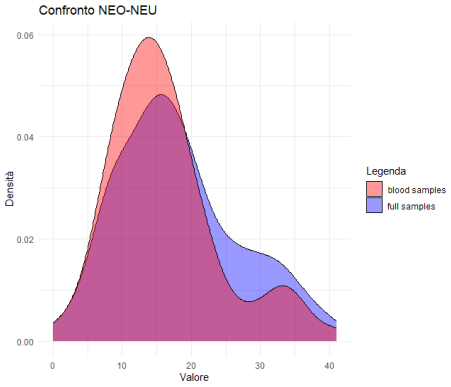

FIG. 12: Confronto NEO NEU campione std con campione blood test

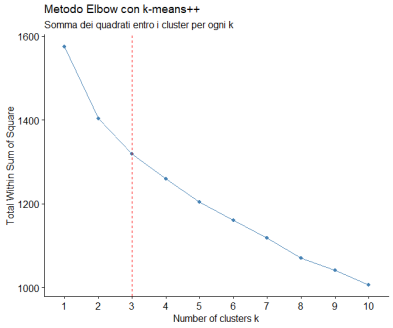

FIG. 13: Metodo elbow

7. DEFINIZIONE ED ANALISI DEI CLUSTER

Per la fase di clustering è stato scelto l'algoritmo **Kmeans**. Per scegliere il numero di cluster ottimale da generare, sono state usate due tecniche.

Elbow method: Il metodo **Elbow** ci fornisce il numero ottimale di cluster calcolando la somma al quadrato delle distanze intra-cluster per i diversi valori di K, costruendo un grafico di questo andamento, si forma un gomito, che indica il valore ottimale di cluster da usare.

$$WCSS(K) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Silhouette method: Il metodo della silhouette è una tecnica di validazione per algoritmi di clustering (come K-means o gerarchico) che misura la qualità dei raggruppamenti basandosi sulla coesione interna (distanza tra punti dello stesso cluster) e sulla separazione (distanza tra cluster diversi). Il coefficiente varia da -1 a +1: valori vicini a +1 indicano cluster ben separati e compatti, mentre valori negativi suggeriscono una cattiva classificazione. Per ogni punto i, il coefficiente $s(i)$ è definito come $\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, dove $a(i)$ è la distanza media dal proprio cluster e $b(i)$ è la distanza media dal cluster più vicino.

7.1. Tuning ottimale

Sia il metodo **Elbow**, sia quello della **Silhouette**, come riportato nelle figure (13 e 14), il numero ottimale di cluster risulta essere tre, anche se il metodo della Silhouette ci indica vista la complessità del cluster, che ci saranno dei confini poco netti fra i cluster, visto l'indice tra i cluster basso. In questo lavoro è stato usato **Kmeans++**, che mitiga alcuni problemi della versione base.

7.2. Metriche di validazione

I Cluster generati(Fig 15) sono stati validati con le seguenti metriche:

- **WCSS** Tramite la metrica della somma intra cluster quadratica possiamo capire la coesione dei cluster.
- **BCSS** Ci aiuta a capire quanto i cluster si distanziano fra di essi.
- **Calinski-Harabasz** è una metrica di valutazione interna utilizzata per determinare la qualità di un algoritmo di clustering.

7.3. Risultati clustering

I Cluster generati(Fig 15) , hanno una struttura molto complessa, vista la complessità del dataset ci aspettavamo che **Kmeans++** facesse difficoltà, ma abbiamo scelto di usare kmeans per testare il suo comportamento su un dataset complesso e confrontarlo con tecniche quantistiche.
Il confronto tra 2 e 3 cluster ha prodotto i risultati riportati nella tabella I.

TABLE I: Confronto delle metriche di clustering (2 vs 3 Cluster)

Metrica	2 Cluster	3 Cluster
WCSS	801.282, 601.699	355.868, 499.591, 471.545
BCSS	773.718, 973.301	1219.132, 1075.409, 1103.455
Calinski-Harabasz	71.454, 119.701	125.042, 78.569, 85.413

Come si evince dalla tabella riportata, il migliore cluster risulta essere il raggruppamento in tre gruppi.

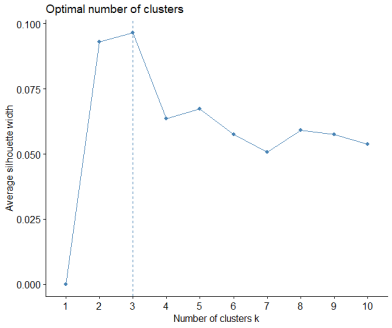

FIG. 14: Metodo Silhouette

7.4. RQ 1: Esiste una correlazione significativa tra una diagnosi di depressione e una diagnosi positiva di demenza?

L'analisi dei cluster suggerisce una correlazione complessa e non lineare, mediata da profili clinici distinti:

Cluster 1: Questo gruppo presenta i livelli più elevati di Depressione (BDI) e di Coping Evitante. Contestualmente, mostra i valori più bassi in Ossigenazione & Immunità e punteggi deficitari nella Memoria (CVLT Dim1). La combinazione di alta depressione e compromissione dei biomarcatori suggerisce che in questo sottogruppo la depressione possa agire come fattore di vulnerabilità o segnale precoce di declino cognitivo.

Cluster 2: Questo gruppo presenta livelli intermedi di depressione, ma le migliori performance in assoluto nella Memoria e nei parametri di Ossigenazione & Immunità. È interessante notare che questo cluster è composto prevalentemente dal sesso maschile.

Cluster 3: Presenta i livelli più bassi di depressione e una riserva cognitiva (RPM) nella media, associata a una performance di memoria superiore al Cluster 1 ma inferiore al Cluster 2.

Alla luce di questi dati, la risposta alla **RQ1** è affermativa ma polarizzata:

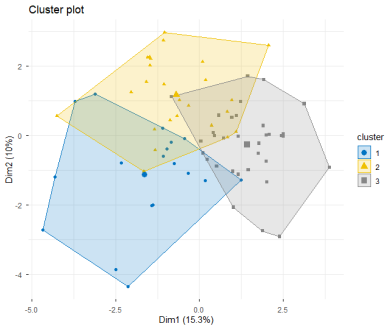

FIG. 15: Cluster ottenuto

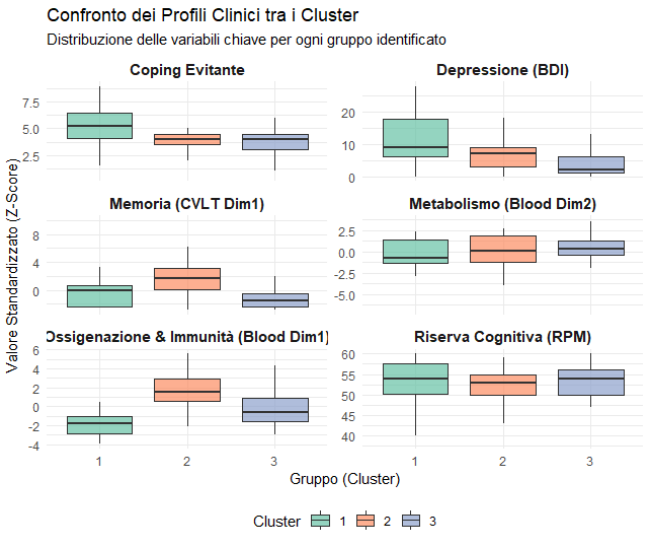

FIG. 16: Risultati dei gruppi di cluster

Esiste una correlazione significativa visibile nel Cluster 1, dove il picco della sintomatologia depressiva coesiste con i deficit di memoria e l'indebolimento del sistema immunitario/ossigenazione.

7.5. RQ 2: È possibile identificare pattern alternativi o non convenzionali utili ai fini della diagnosi precoce?

è possibile identificare pattern non convenzionali utili alla diagnosi precoce, poiché l'analisi dei cluster evidenzia limitazioni nei marcatori di rischio tradizionali e fa emergere configurazioni cliniche alternative.

In particolare emerge dall'osservazione di specifiche

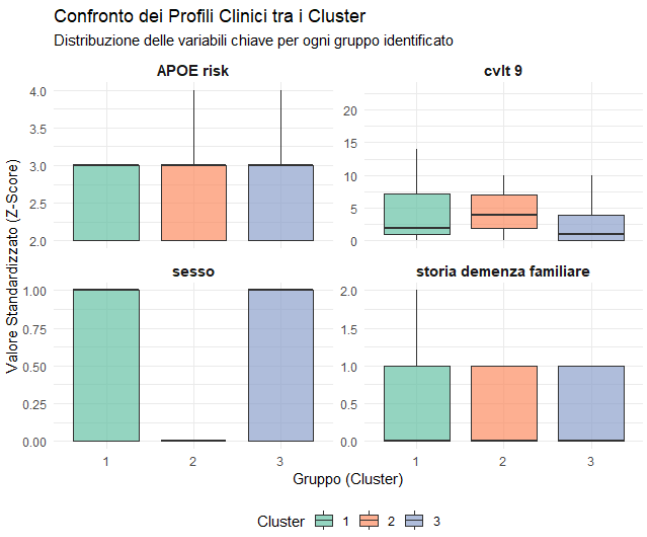

FIG. 17: Risultati dei gruppi di cluster

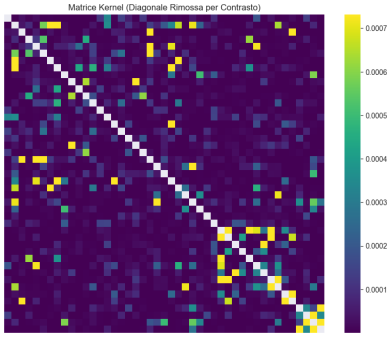

FIG. 18: Kernel mappato su un rotazione massima di pi/2

dissociazioni tra sfere biologiche, psicologiche e cognitive:

Cluster 1: Emerge un pattern in cui il declino è preceduto da alti livelli di depressione e coping evitante, pur in presenza di una riserva cognitiva che maschera i deficit di memoria ai test standard.

Cluster 2: Si identifica un profilo in cui l'integrità della memoria è sostenuta da elevati livelli di ossigenazione e immunità, che agiscono come fattori protettivi indipendenti dalla predisposizione genetica.

Cluster 3: Si rileva un fenotipo caratterizzato da grave compromissione mnestica (bassi punteggi CVLT) che, non associandosi a sintomatologia depressiva, rischia di sfuggire allo screening clinico basato sul disagio riferito dal paziente.

8. QUANTUM

8.1. Analisi dell'Encoding ZZ e delle metodologie di calcolo dei Kernel Quantistici

Nel contesto del machine learning quantistico, la capacità di rappresentare i dati e calcolarne la similarità

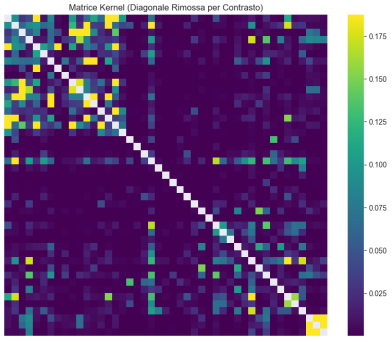

FIG. 19: Kernel mappato su un rotazione massima di 0,5

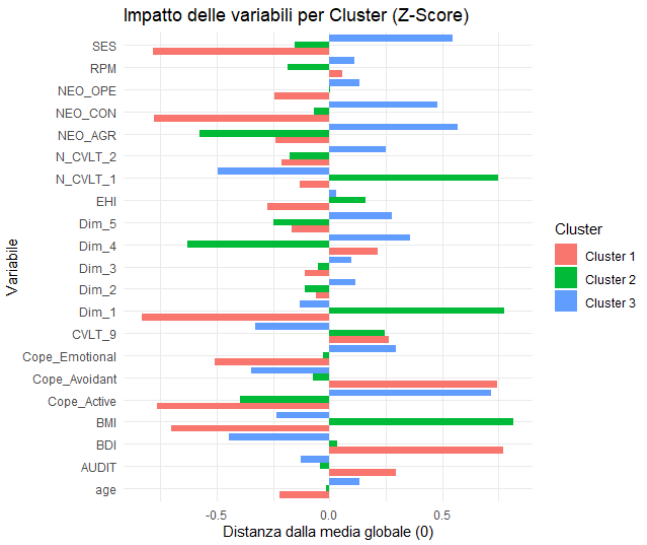

FIG. 20: Features discriminanti per il cluster kmeans classico

è fondamentale. Sulla base delle fonti analizzate, vengono qui descritti il funzionamento della *ZZFeatureMap* e le differenze tra i due principali approcci di calcolo dei kernel: i *Fidelity Quantum Kernels* (FQK) e i *Projected Quantum Kernels* (PQK).

L'Encoding dei dati tramite ZZFeatureMap

La **ZZFeatureMap** è un circuito quantistico parametrizzato di secondo ordine, progettato per codificare dati classici all'interno di uno stato quantistico. Il suo funzionamento si basa sull'applicazione iniziale di porte di Hadamard per generare sovrapposizione, seguite da rotazioni di fase R_Z dipendenti dai dati in ingresso.

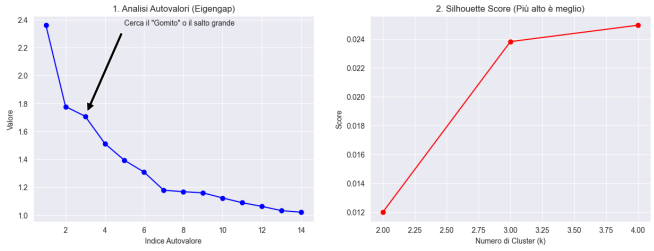

FIG. 21: metriche quantistiche per ottenere il miglior numero di cluster

La peculiarità di questo circuito risiede nell'introduzione di interazioni tra coppie di qubit (le interazioni "ZZ"), che permettono di codificare non solo le singole feature, ma anche le correlazioni esistenti tra di esse attraverso l'*entanglement*. Sebbene sia una tecnica standard, lo studio ha evidenziato che la `ZZFeatureMap` può mostrare performance lim-

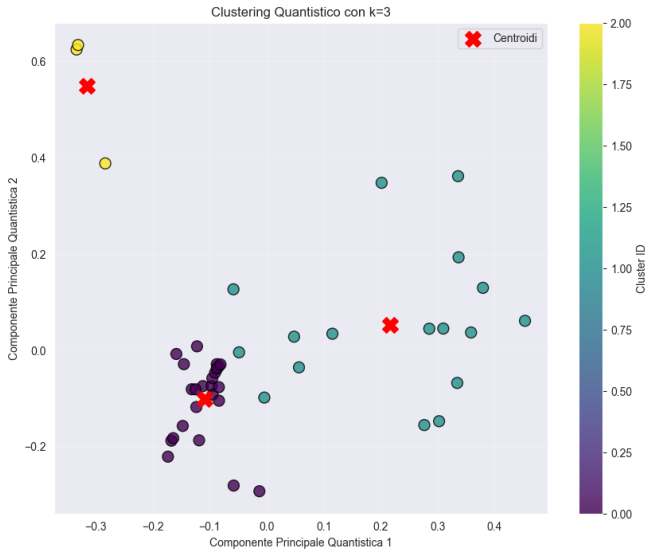

FIG. 22: Cluster usando kernel Fidelity

itate in specifici task di regressione quando valutata singolarmente, pur mantenendo prestazioni aggregate comparabili ad altri circuiti su famiglie di dataset più ampie.

1. *Metodi di calcolo: Fidelity vs Projected Kernel*

Una volta codificati i dati, è necessario definire una misura di similarità (kernel). L'approccio più immediato è rappresentato dal **Fidelity Quantum Kernel (FQK)**, che opera sfruttando la geometria nativa dello spazio di Hilbert. L'FQK calcola la similarità tra due dati x e x' misurando la "fedeltà", ovvero la sovrapposizione tra i rispettivi stati quantistici $|\psi(x,\theta)\rangle$ e $|\psi(x',\theta)\rangle$. Matematicamente, questo corrisponde al modulo quadro del loro prodotto interno:

$$k_\theta^{\mathrm{FQK}}(x,x') = |\langle\psi(x,\theta)|\psi(x',\theta)\rangle|^2 \tag{1}$$

Nonostante la sua eleganza teorica, l'FQK è soggetto al fenomeno della "concentrazione esponenziale": all'aumentare del numero di qubit, i valori del kernel tendono a concentrarsi in un intervallo ristretto, rendendo difficile l'addestramento dei modelli su larga scala.

Per mitigare questa problematica sono stati introdotti i **Projected Quantum Kernels (PQK)**. A differenza dell'approccio precedente, i PQK non utilizzano l'intero stato quantistico, ma proiettano l'informazione su una rappresentazione classica approssimata prima del confronto. Il processo prevede di effettuare misurazioni locali su sottoinsiemi di qubit (ad esempio tramite matrici densità ridotte o *k-RDM*) per estrarre vettori di feature. Questi vettori vengono successivamente elaborati da un kernel classico esterno (detto *outer kernel*), come ad esempio un kernel Gaussiano. Una formulazione tipica del PQK è data da:

$$k_\theta^{\mathrm{PQK}}(x,x') = \exp\left(-\gamma \sum_{k,P} \left[\mathrm{tr}(\rho_\theta(x)P_k) - \mathrm{tr}(\rho_\theta(x')P_k)\right]^2\right) \tag{2}$$

L'efficacia del PQK dipende fortemente dalle scelte progettuali, come il tipo di kernel esterno e gli operatori di misurazione selezionati.

2. *Conclusioni dello studio comparativo*

Contrariamente alle previsioni teoriche che vedevano i PQK favoriti grazie alla loro resistenza alla concentrazione esponenziale, lo studio di benchmarking ha rivelato che, per le dimensioni analizzate (fino a 15 qubit), non vi sono differenze significative di prestazioni tra FQK e PQK. I risultati suggeriscono che il fattore determinante non è tanto la scelta tra i due metodi, quanto un'attenta ottimizzazione degli iperparametri, in particolare la regolarizzazione e il *bandwidth tuning* (γ).

8.2. Pre-Applicazione

La sperimentazione su computer reale è stata effettuata sul computer quantistico di IBM Italia composto da 133 q-bit. In questa fase abbiamo utilizzato solo il 60% del campione:

Cope-Avoidant, Cope_Active, BDI, SES, EHI, NEO_CON, NEO_AGR, AUDIT, Cope_Emotional, CVLT_9, N_CVLT_1, N_CVLT_2, Dim_1, Dim_4

Per osservare quali features sono più influenti nella generazione dei cluster standard. Al fine ultimo di confrontare i cluster *classici* con quelli *quantistici* abbiamo utilizzato le seguenti metriche di confronto:

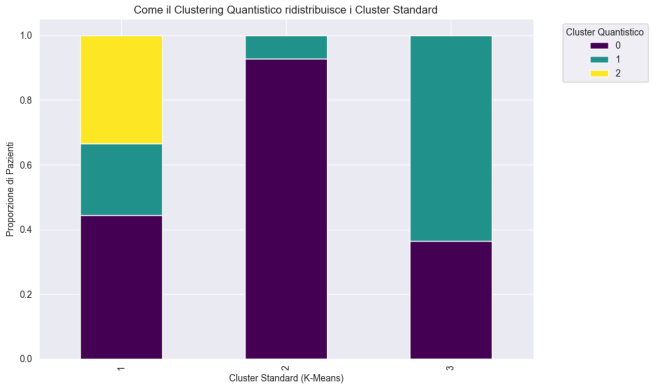

FIG. 23: Confronto cluster ottenuto da R e FQK cluster

TABLE II: Valutazione della Somiglianza dei Raggruppamenti

Metrica	Valore
Somiglianza dei Raggruppamenti (ARI)	0.201
Informazione Condivisa (NMI)	0.275

Adjusted Rand Index: serve a misurare la qualità di un algoritmo di clustering. Confronta i raggruppamenti generati dall'algoritmo con le etichette reali (o con un altro clustering) per valutarne la coerenza. Il suo grande vantaggio è l'indipendenza dai nomi assegnati alle categorie: l'indice riconosce infatti che due gruppi sono identici se contengono gli stessi dati, anche se etichettati diversamente. Infine, restituisce un punteggio normalizzato di facile lettura, dove il valore 1 indica una sovrapposizione perfetta, mentre lo 0 segnala un risultato totalmente casuale.

Normalized Mutual Information: serve a confrontare i gruppi trovati dall'algoritmo con le etichette reali, ma lo fa usando un approccio diverso: la Teoria dell'Informazione. valuta la qualità del clustering basandosi sulla quantità di informazione condivisa. Invece di contare le coppie come l'ARI, si chiede: "Se conosco il gruppo che l'algoritmo ha assegnato a un dato, quanto diventa più facile indovinarne la vera etichetta?". Il risultato viene poi normalizzato tra 0 (nessuna informazione condivisa, clustering inutile) e 1 (informazione completa, clustering perfetto), rendendo il confronto affidabile anche quando il numero di cluster e classi non coincide perfettamente.

Spectral Clustering: serve a raggruppare dati che hanno forme complesse e irregolari, dove gli algoritmi classici falliscono. Mentre metodi come il K-Means assumono che i gruppi siano compatti e sferici, lo Spectral Clustering approccia il problema dal punto di vista della connettività: interpreta i dati come una rete interconnessa e cerca di "tagliare" i collegamenti più deboli. Questo lo rende indispensabile in scenari complessi come la segmentazione delle immagini o l'analisi di dati con strutture non lineari, permettendo di distinguere correttamente cluster intrecciati o annidati che risulterebbero indistinguibili per altri metodi.

8.3. Applicazione Simulata

Seguendo i consigli del paper[5] da cui abbiamo attinto le informazioni, siamo partiti utilizzando la **ZZFeatureMap** per poter codificare il nostro dataset su un computer quantistico. In questo caso, mappato nello spazio di *Hilbert* ad altà dimensionalità, codificando nelle rotazioni dei qubit usando un entanglement leggero.

1. *Fidelity Quantum Kernel*

Siamo partiti utilizzando il Fidelity Quantum Kernel (FQK) che come riportato sfrutta la geometria dello spazio di hilbert che verificare la somiglianza fra elementi. I parametri fondamentali del kernel preso in esame è la codifica di questo spazio, il paper suggeriva di partire con un range $[0, \pi]$, ma noi siamo partiti con uno spazio più ristretto $[0, \pi/2]$. Come riportato nella figura 18 su una scala da 0 a 1, i range di questo kernel risultano essere molto al di sotto per la costruzione di cluster.

Per questo motivo abbiamo abbiamo ristretto la rotazione in una mappatura $[0, 0.5]$, ottenendo un incremento della correlazione fra le entità, avendoli ristretti nello spazio. Come illustra la figura 19, il kernel ha catturato meglio la somiglianza dei campioni ma non abbastanza, quindi ci si trova davanti a quello che si chiama "*Exponential Concentration*" l'algoritmo pensa che tutti i punti siano dissimili non riuscendo a catturare bene le relazioni, per questo motivo non siamo andati oltre e abbiamo costruito i cluster sui dati ottenuti, per confrontarli e successivamente implementare la soluzione **PQK**.
Applicando le misura che corrispondo ad elbow e silhouette quantistiche, anche qui otteniamo 3 come miglior numero di cluster. La figura 22 che rappresenta il cluster ottenuti tramite SpectralCluster, ci mostra come il primo gruppo sia isolato e di pochi campioni, il gruppo 2 in viola è molto denso e vicino ai bordi dell'ultimo gruppo che risulta essere più dispersivo. Si è poi confrontato, attraverso le metriche *Adjusted Rand Index* e *Normalized Mutual Information* , i risultati del cluster generato, quanto combaciassero con il cluster ottenuto con i metodi tradizionali riportati nella tabella II, i due cluster sono molto differenti classificando elementi in modi diversi.

Come si può vedere anche tramite la Fig. 23. solo il

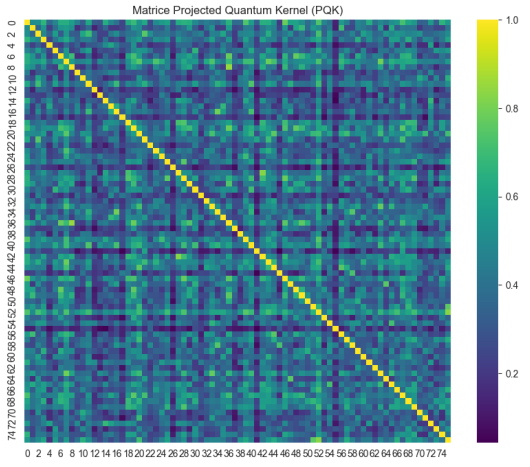

FIG. 24: Kernel Projected quantum kernel

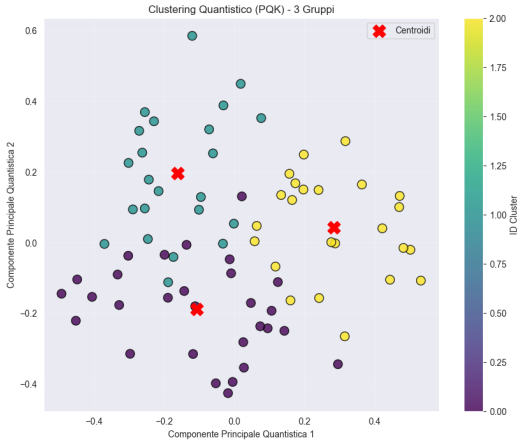

FIG. 25: Cluster pqk ottenuto

cluster due risulta essere coincidente con il primo gruppo ottenuto dalle tecniche standard, ribadendo ciò che è stato detto in precedenza.

2. *Projected Quantum Kernel*

Per cercare di mitigare l' *"Exponential Concentration"* si è seguita la raccomandazione del paper analizzato e quindi abbiamo applicato il PQK. Essendo questo metodo ha una complessità temporale minore, abbiamo riportato la grandezza del campione alle 76 osservazioni finali.
Attraverso le spiegazioni e di gemini abbiamo applicato una correzione automatica alla gamma che influisce sullo "zoom" dei dati, tarando correttamente i valori del kernel. Come si può vedere dalla Fig. 24, questo approccio cattura maggiormente le relazioni fra i dati, ottenendo un netto miglioramento rispetto l'approccio basato sul Fidelity Quantum Kernel. Si è quindi proceduto con la creazione dei cluster tramite SpectralCluster, ottenendo ciò che è riportato qui 25. La struttura del cluster generato sembra molto più ben strutturata rispetto all'approccio precedente, andando a distinguere 3 gruppi ben definiti anche se sovrapposti ai bordi dei cluster. An-

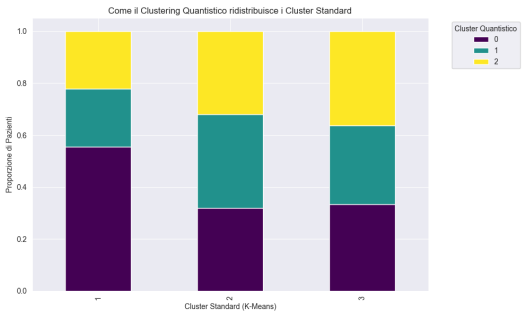

FIG. 26: Comparazione del cluster ottenuto in R con il cluster PQK simulato

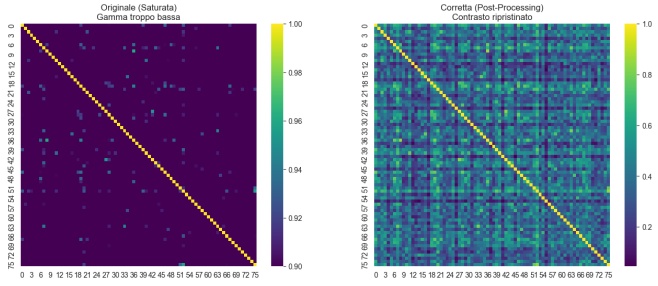

FIG. 27: Correzione di Gamma PQK sui dati elaborati sul un QPU reale

TABLE III: Valutazione delle prestazioni del clustering PQK

Metrica	**Valore**
Somiglianza dei Raggruppamenti (ARI)	−0.011
Informazione Condivisa (NMI)	0.019

che questo clustering come riportato nella tabella III e in Figura 26 risulta essere distante dai gruppi definiti con l'approccio standard se non di più rispetto al FQK. Le metriche ci indicano che la correlazione fra le due soluzione è praticamente nulla, come si evince anche dal grafico riportato.

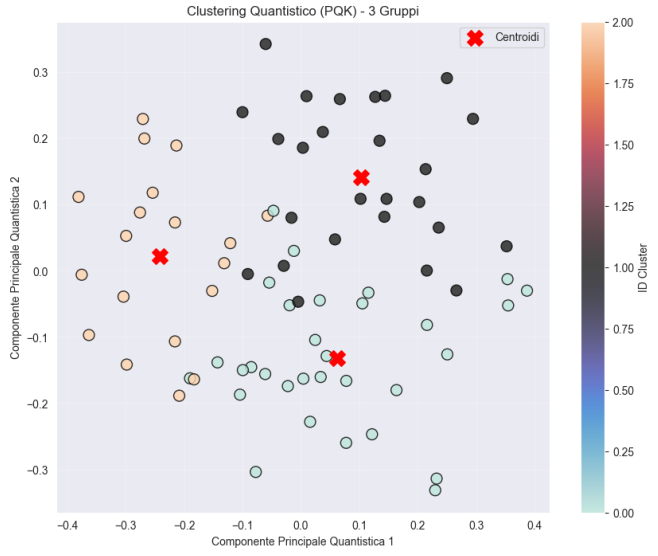

FIG. 28: Spectral cluster ottenuto da PQK su esecuzione reale IBM torino

TABLE IV: Valutazione della Somiglianza dei Raggruppamenti

Metrica	Valore
Somiglianza dei Raggruppamenti (ARI)	0.295
Informazione Condivisa (NMI)	0.337

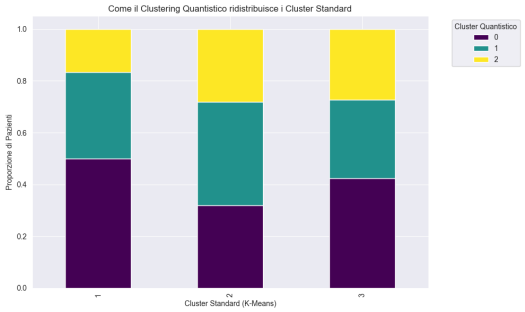

FIG. 29: Cluster pqk ottenuto

3. Specifiche del computer quantistico reale

8.4. Applicazione su computer quantistico reale

Per motivi di rumore quantistico, le features usando il FQK si sovrappongono andando a compromettere i dati e la mappatura, su tutti e 3 i livelli di codifica indicati da IBM. Per tanto, l'esecuzione sul computer quantistico di torino IBM è avvenuta solamente usando l'approccio PQK.
Per poter eseguire, quindi, solamente il codice del PQK è stata necessaria una fase di transpilazione[1], cosi da poter codificare la nostra simulazione nelle istruzioni base per poter essere eseguito sul computer quantistico reale.

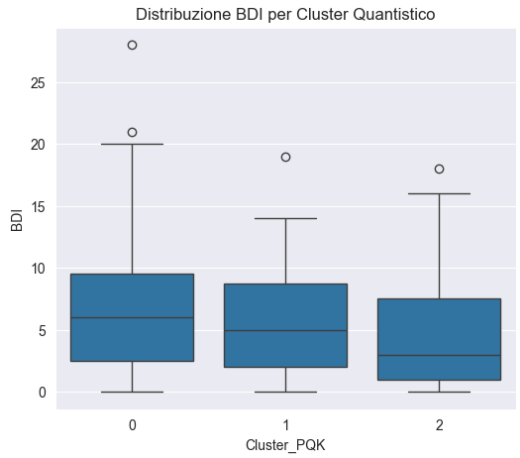

FIG. 30: Boxplot distribuzione BDI del cluster quantistico con PQK

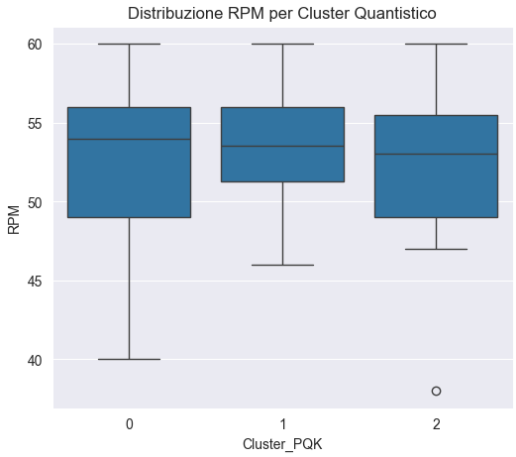

FIG. 31: Cluster pqk ottenuto

I risultati ottenuti sono salvati nella repository del progetto. Una volta ottenuta la matrice di kernel, essa risulta fuori gamma usando il metodo di gauss, rischiando di compromettere il lavoro eseguito.Tramite gemini è stato possibile ripristinare i valori della matrice ed applicare la gamma ottimale, come mostrato in Fig. 27, risultando molto simile alla gamma della simulazione. Avendo corretto la gamma si sono generati i cluster tramite SpectralCluster, riportati nella Fig. 28, questa struttura presenta una topologia strutturale equivalente a quella della simulazione illustrata in Fig. 25, effettuando una rotazione di 90 verso sinistra del cluster generato dal kernel ottenuto dal computer IBM, si può notare come alcuni punti del cluster e della struttura si sovrappongono, come indicato anche nelle metriche riportate nella tabella IV. Suggerendo una distorsione per via del rumore e computazione su hardware reale del calcolo degli autovalori, un dato inaspettato che riportiamo per coerenza dei risultati tra una simulazione e ciò che si otterrebbe realmente.

In ultima analisi, vista la topologia condivisa con la simulazione anche il risultato dell'esecuzione reale risulta generare un cluster totalmente differente dagli approcci std, come si nota dalla Fig. 29

[1] La transpilazione è il processo di riscrittura di un determinato circuito di input per adattarlo alla topologia di uno specifico dispositivo quantistico e ottimizzare le istruzioni del circuito per l'esecuzione su computer quantistici rumorosi.

8.5. RQ3 3:Quali differenze emergono dal confronto tra i risultati ottenuti mediante tecniche di clustering tradizionale e quelli derivanti da algoritmi di clustering quantistico?

Come mostrato nella sezione precedente, la struttura dei cluster risulta essere molto differente, con somiglianze puramente casuali. I due approcci hanno creato gruppi differenti, che potrebbe fornire una visione diversa ma costruttiva del progetto in esame. L'approccio PQK è risultato essere quello più robusto, rispetto FQK. Per concludere, il differenze sono nette ma anche con un approccio quantistico è emersa una correlazione fra la depressione e riserva cognitiva, analizzando la Fig. 30 e 31, si nota come nel primo gruppo vi sia una lunga coda rappresentante una forte depressione associata ad un indice RPM molto basso che rappresenta la riserva cognitiva di una persona e quindi un fattore di rischio d'insorgenza di demenza grave e di sfociare in Alzheimer. Quindi si, le differenze negli approcci ci sono state ma sono sorti dei pattern comuni su alcuni gruppi.

9. CONCLUSIONI

Il presente studio ha indagato la diagnosi precoce dell'Alzheimer integrando analisi statistica tradizionale e Quantum Machine Learning su un campione di pazienti di mezza età.

Dal punto di vista clinico, i risultati confermano una correlazione complessa e non lineare tra depressione e declino cognitivo (RQ1), isolando specifici fenotipi a rischio spesso mascherati dalla "riserva cognitiva" (RQ2).

Sul piano metodologico (RQ3), il confronto tra approcci classici e quantistici ha evidenziato profonde divergenze strutturali. L'algoritmo Projected Quantum Kernel (PQK) si è dimostrato superiore al Fidelity Quantum Kernel nel mitigare la concentrazione dei dati, generando raggruppamenti inediti e distinti rispetto al clustering classico. L'esecuzione con successo su hardware reale IBM-Torino (133 qubit) ha validato la coerenza topologica delle simulazioni, suggerendo che il calcolo quantistico possa offrire prospettive complementari fondamentali per decifrare la complessità eterogenea delle patologie neurodegenerative.

10. BIBLIOGRAFIA

[1] Kunie Ando, Siranjeevi Nagaraj, Fahri Küçükali, Marie-Ange De Fisenne, Andreea-Claudia Kosa, Emilie Doeraene, Lidia Lopez Gutierrez, Jean-Pierre Brion, and Karelle Leroy. Picalm and alzheimer's disease: an update and perspectives. *Cells*, 2022.

[2] P. Dzianok and E. Kublik. PEARL-Neuro database: EEG, fMRI, health and lifestyle data of middle-aged people at risk of dementia. *Scientific Data*, 2024.

[3] C. A. Lane, J. Hardy, and J. M. Schott. Alzheimer's disease. *European Journal of Neurology*, 2018.

[4] Ana-Caroline Raulin, Sydney V Doss, Zachary A Trottier, Tadafumi C Ikezu, Guojun Bu, and Chia-Chen Liu. Apoe in alzheimer's disease: pathophysiology and therapeutic strategies. *Molecular neurodegeneration*, 2022.

[5] Jan Schnabel and Marco Roth. Quantum kernel methods under scrutiny: a benchmarking study. *Quantum Machine Intelligence*, 2025.