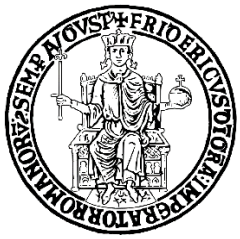
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II



NATIONAL PH.D. PROGRAM IN
ARTIFICIAL INTELLIGENCE

THESIS FOR THE DEGREE OF DOCTOR IN PHILOSOPHY

Definizione di profili metabolici per la certificazione di qualità e di origine di prodotti agroalimentari e loro tracciabilità attraverso intelligenza artificiale e tecnologia BlockChain

Supervisor
Prof. Riccardo SPACCINI

Candidate
Silvana CANGEMI

DR995914

Co-Supervisor
Dott. Attilio MONDELLI

37th Cycle

Indice

Lista delle Figure

Lista delle Tabelle

Capitolo 1 Introduzione

1.1.1 Contesto e motivazione dello studio

L'istituzione dei sistemi di certificazione di origine e qualità risponde all'esigenza crescente di tutelare i consumatori, valorizzare le produzioni agroalimentari tipiche e proteggere i produttori da pratiche concorrenziali scorrette. In un contesto sempre più influenzato dalla globalizzazione dei mercati, dall'allungamento delle filiere produttive e dal progressivo aumento delle sofisticazioni fraudolente, tali sistemi si configurano come strumenti formali in grado di garantire, in modo oggettivo e verificabile, l'autenticità, la provenienza e la qualità dei prodotti alimentari.

Tra i principali strumenti riconosciuti a livello europeo si annoverano la Denominazione di Origine Protetta (DOP), l'Indicazione Geografica Protetta (IGP) e la Specialità Tradizionale Garantita (STG), disciplinati dal Regolamento (UE) n. 1151/2012. Queste certificazioni si fondano sul riconoscimento del legame tra il prodotto e il territorio di origine, promuovendo la trasparenza informativa e offrendo garanzie concrete sia ai consumatori che ai produttori (Chilla et al., 2020; Adinolfi et al., 2024). Inoltre, esse rappresentano un meccanismo strategico per sostenere le economie locali, incentivare pratiche agricole sostenibili e preservare la biodiversità e il patrimonio culturale legato alle tradizioni agroalimentari (Biénabe & Marie-Vivien, 2017; Cei, Defrancesco, & Stefani, 2018).

Affinché tali sistemi risultino effettivi e credibili, è indispensabile disporre di sistemi di tracciabilità affidabili, in grado di documentare ogni fase della filiera agroalimentare, dalla produzione primaria fino alla trasformazione, distribuzione e consumo. La tracciabilità costituisce infatti la base operativa e informativa su cui si fonda l'intero impianto certificativo, poiché consente di seguire la storia del prodotto, verificarne la conformità ai disciplinari e garantire la trasparenza nei confronti di consumatori, enti di controllo e operatori economici. In questo senso, certificazione e tracciabilità agiscono in stretta sinergia: la prima formalizza e attesta il valore del prodotto, mentre la seconda ne fornisce le evidenze oggettive.

L'efficacia di tale interazione è stata progressivamente rafforzata grazie all'evoluzione tecnologica, che ha portato al superamento dei tradizionali sistemi cartacei a favore di soluzioni digitalizzate. Queste innovazioni hanno migliorato in modo significativo la capacità di verifica, la sicurezza nella gestione dei dati e l'affidabilità complessiva del processo di certificazione.

Come mostrato nella Tabella 1.1, gli sviluppi più recenti includono l'introduzione di sensori IoT per il monitoraggio in tempo reale delle condizioni produttive e logistiche, e l'impiego della

blockchain per la registrazione sicura, immutabile e decentralizzata delle informazioni lungo la filiera agroalimentare (Patelli & Mandrioli, 2020; Gálvez, Mejuto, & Simal-Gándara, 2018).

Tabella 1.1 - Evoluzione dei sistemi di tracciabilità: confronto tra strumenti tradizionali, digitali e distribuiti, con evidenza di punti di forza e criticità operative.

Tecnologia di tracciabilità	**Esempi pratici**	**Vantaggi principali**	**Limiti e criticità**
Cartacea / documentale	Registri di produzione, etichette manuali, certificazioni cartacee	Basso costo iniziale; semplicità di utilizzo	Elevata soggettività; rischio di errori umani o frodi; difficile verifica
Digitale centralizzata	Codici a barre, QR code, sistemi ERP, banche dati aziendali	Automazione del processo; accesso rapido ai dati	Dipendenza da server o database centralizzati; rischio di perdita o manipolazione dei dati; scarsa interoperabilità
Sensoristica + tracciabilità distribuita (IoT + Blockchain)	Sensori georeferenziati, RFID, smart contracts	Sicurezza e immutabilità dei dati; trasparenza lungo la filiera; verifica decentralizzata	Costi iniziali elevati; necessità di infrastrutture tecnologiche; standardizzazione ancora in evoluzione; competenze tecniche richieste

Tuttavia, una delle principali sfide nella costruzione di sistemi affidabili di certificazione e tracciabilità rimane la valutazione oggettiva delle caratteristiche qualitative degli alimenti, fondamentale anche per l'effettuazione di controlli di autenticità lungo la filiera.

Un approccio tradizionale a tale scopo consiste nel rilevare variazioni nella composizione metabolica dei prodotti attraverso metodi analitici *targeted*, che prevedono la ricerca e, se necessario, la quantificazione di un numero limitato di composti noti *a priori* (ad esempio adulteranti o contaminanti specifici). Sebbene tali metodi siano consolidati e ben validati, possono risultare inefficaci o insufficienti nei casi in cui la qualità o l'autenticità di un alimento sia influenzata da sostanze non note, la cui presenza può indicare alterazioni, sofisticazioni o frodi accidentali o intenzionali.

In questo contesto, gli approcci non-targeted stanno assumendo un ruolo sempre più rilevante nel controllo qualità e nella verifica dell'autenticità, poiché permettono di ottenere una visione globale e priva di ipotesi analitiche predefinite sulla composizione chimica della matrice alimentare.

Tra le tecniche emergenti, l'impiego della spettroscopia NMR (Risonanza Magnetica Nucleare) si è affermato come uno strumento particolarmente efficace per la caratterizzazione metabolomica dei prodotti agroalimentari. Essa consente di acquisire in modo rapido,

riproducibile e non distruttivo il fingerprint metabolico completo di un campione, generando un'impronta digitale unica, rappresentativa dell'origine, della qualità e degli eventuali trattamenti subiti dal prodotto (Emwas et al., 2019).

L'impiego di algoritmi di intelligenza artificiale (AI) applicati ai dati spettrali generati dalla spettroscopia NMR consente di automatizzare l'analisi dei risultati, individuando correlazioni complesse tra il profilo metabolico e variabili di interesse quali l'origine geografica, la varietà, lo stato di trasformazione o la presenza di adulterazioni. Questi strumenti permettono la costruzione di modelli predittivi robusti e generalizzabili, capaci di classificare accuratamente i campioni e di supportare la tracciabilità lungo l'intera filiera produttiva.

L'integrazione con architetture digitali avanzate, che includono sistemi blockchain e dispositivi IoT per la registrazione sicura e immutabile dei dati, consente lo sviluppo di sistemi di certificazione innovativi, fondati su informazioni oggettive, scientificamente validate e resistenti alla manomissione.

Questa architettura integrata NMR–AI–Blockchain non solo accresce l'efficacia nel rilevare prodotti non conformi o fraudolenti, ma fornisce anche strumenti di verifica indipendenti, accessibili in tempo reale da parte di enti certificatori, produttori e consumatori. In tal modo, il sistema non si limita alla tutela della qualità e dell'autenticità, ma si configura come un mezzo di valorizzazione strategica delle produzioni certificate, contribuendo a rafforzare la competitività delle filiere agroalimentari di eccellenza.

1.1.2 Obiettivo del lavoro

Il presente lavoro di tesi ha avuto come obiettivo lo sviluppo e l'ottimizzazione di protocolli analitici basati su tecniche di machine learning per la gestione di profili metabolici ottenuti mediante spettroscopia NMR. L'attività è stata finalizzata alla costruzione di *biofingerprint* affidabili, riproducibili e scientificamente validati, rappresentando il primo passo verso la loro integrazione in sistemi blockchain per la tracciabilità dei prodotti agroalimentari. L'attività svolta si inserisce all'interno di un progetto di ricerca nato dalla collaborazione con Farzati S.p.A., realtà di eccellenza nel panorama dell'innovazione applicata ai sistemi di certificazione, tracciabilità e valorizzazione delle produzioni agroalimentari.

Si è inteso realizzare un flusso di lavoro capace di processare e interpretare i dati metabolomici in modo efficiente e standardizzato, riducendo al minimo la soggettività dell'operatore, ottimizzando i tempi di analisi e garantendo la possibilità di gestire elevati volumi di dati. L'acquisizione dei profili spettrali è stata effettuata avvalendosi delle tecniche HR-MAS e CP-

MAS NMR, approcci che consentono l'analisi diretta dei campioni senza necessità di estrazioni, preservando la rappresentatività chimico-fisica della matrice originale.
Per assicurare la robustezza metodologica e la trasferibilità dei protocolli, in vista di un'applicazione su larga scala lungo la filiera produttiva, sono state selezionate matrici differenti, rappresentative di diversi livelli di trasformazione e caratterizzate da una natura chimico-fisica eterogenea. Tale approccio ha consentito di verificare l'efficacia dei protocolli sviluppati su tipologie di campioni tra loro molto diverse. Il frumento è stato utilizzato come esempio di materia prima allo stato naturale, priva di significativi processi di trasformazione industriale; il latte come prodotto intermedio, sottoposto a un primo trattamento tecnologico ma ancora chimicamente prossimo alla materia prima; e i concentrati di pomodoro come rappresentazione di un alimento fortemente trasformato e complesso. Questa impostazione ha consentito di esplorare le potenzialità e i limiti dei protocolli in contesti applicativi differenti, simulando criticità reali lungo le diverse fasi della filiera, dal campo al prodotto finito.
Contestualmente, è stata avviata una linea di ricerca finalizzata all'analisi di campioni di cippati vegetali, georeferenziati e forniti dall'azienda Farzati S.p.A., già associati a sistemi di tracciabilità basati su tecnologia IoT e blockchain. L'attività ha previsto l'acquisizione di spettri CP-MAS NMR e la successiva elaborazione chemiometrica, con l'obiettivo di esplorare l'integrazione degli approcci analitici NMR con le infrastrutture di tracciabilità esistenti.
I risultati ottenuti nell'ambito del presente lavoro consolidano le prospettive di applicazione di protocolli analitici avanzati, basati sull'integrazione di spettroscopia NMR, tecniche di machine learning e sistemi di tracciabilità blockchain, per il controllo della qualità e dell'origine dei prodotti agroalimentari. L'attività svolta contribuisce inoltre a favorire sinergie con realtà operative di riferimento, quali Farzati S.p.A., nel contesto dell'innovazione dei sistemi di certificazione e tracciabilità.

1.1.3 Struttura della Tesi

Di seguito viene delineata la struttura con cui è stato organizzato il presente lavoro.

- **Capitolo 1: Introduzione**
  Si presenta il contesto generale della ricerca, con la definizione degli obiettivi principali e si fornisce una panoramica della struttura della tesi.
- **Capitolo 2: Stato dell'Arte e Sfide Tecnologiche**
  Vengono analizzati i principali sviluppi in ambito agroalimentare riguardanti lo studio dei profili metabolici, le tecniche di spettroscopia NMR, gli algoritmi di Intelligenza

Artificiale per il processamento dei dati metabolomici. Vengono inoltre esplorate le possibili integrazioni in sistemi blockchain, evidenziando le sfide tecniche e le limitazioni esistenti.

- **Capitolo 3: Materiali e Metodi**

  Vengono illustrate le procedure sperimentali adottate per l'analisi delle matrici, lo sviluppo degli script computazionali e l'applicazione di metodologie statistiche e algoritmi di Intelligenza Artificiale per il trattamento e l'interpretazione dei dati.
- **Capitolo 4: Risultati e Discussione**

  Sono presentati e discussi i risultati ottenuti mediante approcci chemometrici tradizionali e strategie automatizzate basate su Intelligenza Artificiale. Viene inoltre discusso un esempio di integrazione tra analisi spettroscopica e dati acquisiti tramite sensori IoT, finalizzata alla registrazione in blockchain a supporto della tracciabilità e della certificazione digitale delle filiere agricole.
- **Capitolo 5: Conclusioni e Prospettive Future**

  Vengono sintetizzati i principali risultati, evidenziando il contributo della ricerca e proponendo possibili sviluppi futuri.
- **Capitolo 6: Bibliografia**

  È riportato l'elenco delle fonti bibliografiche consultate e citate.
- **Capitolo 7: Appendice**

  Sono riportati estratti significativi del codice sviluppato per il processamento e l'analisi automatizzata dei dati

Capitolo 2 Stato dell'arte e Sfide Tecnologiche

2.1 La Definizione dei Profili Metabolici per garantire origine e qualità dei prodotti agroalimentari

2.1.1 Premessa

Nel settore agroalimentare, garantire l'origine e la qualità dei prodotti è fondamentale per tutelare i consumatori e valorizzare le eccellenze territoriali. Le certificazioni di origine, come la Denominazione di Origine Protetta (DOP) e l'Indicazione Geografica Protetta (IGP), attestano ufficialmente il legame di un prodotto con un territorio specifico e con metodi tradizionali di produzione (Cassago et al., 2021). Tali indicazioni geografiche assicurano che le caratteristiche distintive di un alimento derivino direttamente dal suo luogo di origine e dal know-how locale. Tuttavia, in un contesto di mercati globalizzati, il rischio di frodi alimentari e adulterazioni è elevato: ingredienti meno pregiati possono essere aggiunti o sostituiti fraudolentemente, compromettendo così l'autenticità del prodotto e ingannando il consumatore. È quindi necessario disporre di metodi analitici affidabili per verificare che quanto dichiarato in etichetta corrisponda effettivamente alla composizione reale e alla provenienza geografica dell'alimento (Selamat et al., 2021).

In questo scenario, la metabolomica si è affermata come uno strumento scientifico efficace per l'autenticazione alimentare, poiché permette di ottenere una vera e propria "impronta digitale" (biofingerprint) chimica di un prodotto, ovvero un profilo completo dei metaboliti che lo caratterizzano. Confrontando tali profili metabolici, è possibile differenziare i prodotti in base alla loro origine geografica, identificare eventuali adulterazioni o contaminazioni e fornire evidenze scientifiche a supporto delle certificazioni di qualità (Cassago et al., 2021).

2.1.2 Analisi mirata e non-mirata

La metabolomica è la disciplina che studia in modo estensivo il metaboloma, ovvero l'insieme complessivo delle piccole molecole (metaboliti) presenti in un organismo o in un campione biologico in un determinato momento, e che rappresentano i prodotti finali delle reazioni biochimiche (Wishart, 2008). A differenza di altre discipline "-omiche", quali la genomica, utile per analizzare il patrimonio genetico, o la proteomica, che si concentra sulle proteine, la metabolomica indaga molecole come zuccheri, amminoacidi, acidi organici, polifenoli e lipidi, direttamente influenzate dall'attività cellulare e fortemente condizionate da fattori ambientali,

fornendo una fotografia completa dello stato fisiologico di un organismo o di un alimento (Wishart, 2008).

In ambito agroalimentare, ciò significa che il profilo metabolico di un prodotto riflette non solo la sua composizione naturale, ma anche l'ambiente di coltivazione o allevamento, le caratteristiche del suolo, le condizioni climatiche, la varietà genetica, i trattamenti tecnologici subiti (come essiccazione, fermentazione, pastorizzazione) e le pratiche agricole o industriali adottate (Cevallos-Cevallos et al., 2009; Cubero-Leon et al., 2014). Di conseguenza, i metaboliti possono essere considerati veri e propri marcatori dell'origine geografica e del processo produttivo, rendendo la metabolomica uno strumento estremamente potente per l'autenticazione alimentare e il supporto alle certificazioni di origine (Cubero-Leon et al., 2014).

L'analisi metabolomica viene generalmente condotta secondo due approcci distinti: l'approccio "mirato" (*targeted*), che punta alla quantificazione di un insieme definito di metaboliti noti, e l'approccio "non mirato" (*untargeted*), volto invece a rilevare il profilo metabolico completo del campione analizzato (Cevallos-Cevallos et al., 2009).

In tema di autenticità alimentare, intesa in senso ampio come corretta identificazione della varietà, dell'origine geografica, del sistema produttivo, della lavorazione o dell'eventuale adulterazione, ciascuno di questi approcci presenta vantaggi e limiti specifici. L'approccio mirato può essere più rapido e preciso nell'individuare specifici marker indicativi di adulterazioni già note, ma potrebbe non essere in grado di evidenziare sofisticazioni emergenti o la presenza di additivi sconosciuti. Al contrario, l'approccio non mirato permette di rilevare composti inattesi, risultando quindi particolarmente efficace nella gestione di frodi emergenti (Mialon et al., 2023).

Le tecniche non mirate consentono di effettuare confronti tra campioni differenti attraverso la ricostruzione della cosiddetta "*bio-fingerprint*", ovvero un'impronta chimica complessiva che rappresenta l'identità metabolica del campione in una determinata condizione. Il *fingerprinting metabolico* può essere impiegato come riferimento per distinguere prodotti autentici da quelli contraffatti, offrendo una visione integrata delle variazioni indotte da fattori genetici, ambientali o tecnologici. In particolare, esso risulta estremamente efficace nella classificazione basata sull'origine geografica, data la complessità associata all'effetto del *terroir* e alle numerose variabili ambientali che influenzano il profilo metabolico dei prodotti.

In molti casi, inoltre, un approccio non mirato può includere una fase successiva di identificazione dei composti discriminanti che, pur non essendo necessaria ai fini della

classificazione, può rafforzare le conclusioni ottenute e, in alcuni contesti, orientare lo sviluppo di metodi mirati (Mialon et al., 2023).

2.1.3 Principali tecniche in metabolomica alimentare

La metabolomica alimentare si avvale di tecniche analitiche avanzate in grado di rilevare, identificare e quantificare un'ampia gamma di metaboliti presenti nelle matrici agroalimentari. Le principali metodologie impiegate sono la spettrometria di massa (MS), generalmente accoppiata a cromatografia liquida (LC-MS) o gassosa (GC-MS), e la spettroscopia di risonanza magnetica nucleare (NMR) (Wishart, 2008; Cevallos-Cevallos et al., 2009).

La LC-MS è particolarmente adatta per l'analisi di composti polari e termolabili, grazie alla sua elevata sensibilità e capacità di separazione, mentre la GC-MS trova impiego nell'identificazione di metaboliti volatili e semivolatili, come aromi e componenti lipidici (Cubero-Leon et al., 2014). L'uso di strumenti ad alta risoluzione (HRMS), come gli analizzatori Orbitrap o TOF, consente inoltre una caratterizzazione precisa anche di metaboliti sconosciuti (Selamat et al., 2021).

La Spettroscopia NMR consente sia la quantificazione diretta dei metaboliti che la costruzione di profili metabolici completi, e si configura come una piattaforma ideale per studi di fingerprinting chimico e tracciabilità dell'origine (Wishart, 2008; Selamat et al., 2021).

Oltre a queste, trovano impiego anche tecniche spettroscopiche vibrazionali come la FT-IR (Fourier Transform Infrared), la MIR (Mid-Infrared Spectroscopy), la NIR (Near-Infrared Spectroscopy) e la spettroscopia Raman, spesso utilizzate nell'ambito del controllo qualità e dell'autenticazione alimentare, grazie alla possibilità di acquisire impronte spettrali rapide e non distruttive, sebbene con contenuto informativo meno dettagliato rispetto a MS e NMR (Cevallos-Cevallos et al., 2009; Ellis et al., 2012).

2.1.4 Impronte Metaboliche e Origine Geografica

Il profilo metabolico di un alimento è influenzato da numerosi fattori, quali il genotipo (specie o varietà utilizzata), le condizioni pedoclimatiche (suolo, clima, altitudine) del territorio di coltivazione o allevamento, le pratiche agronomiche o zootecniche, nonché i processi di trasformazione tecnologica e di stagionatura. Questo insieme di elementi, comunemente identificato con il termine francese "terroir", lascia un'impronta chimica caratteristica nel prodotto finale, che può essere rilevata attraverso la ricostruzione del profilo metabolico (Cevallos-Cevallos et al., 2009; Cubero-Leon et al., 2014).

Numerosi studi hanno dimostrato l'efficacia della metabolomica nell'autenticazione dell'origine geografica dei prodotti agroalimentari. In particolare, questa disciplina consente di rilevare le variazioni chimiche indotte dal terroir, rivelandosi utile sia per i prodotti certificati (DOP/IGP), sia per quelli privi di certificazione. La metabolomica è stata applicata con successo a diverse categorie alimentari, tra cui vino, olio, formaggi, frutta, ortaggi e spezie, evidenziando come il profilo metabolico possa distinguere prodotti simili provenienti da zone diverse, anche all'interno della stessa area certificata (Cassago et al., 2021).
Tra i principali filoni di ricerca sviluppatisi negli ultimi anni si possono citare:

- la classificazione tra aree nazionali e internazionali,
- l'identificazione di firme chimiche nei prodotti a denominazione di origine,
- il confronto tra regioni diverse all'interno dello stesso Paese,
- l'analisi delle differenze tra siti produttivi all'interno della stessa denominazione,
- il confronto tra prodotti certificati e non certificati,
- l'impiego della firma metabolica per supportare la registrazione di nuove DOP/IGP.

Nel complesso, la metabolomica si conferma una tecnologia potente, oggettiva e riproducibile per l'autenticazione dell'origine geografica degli alimenti. Essa è in grado di integrare e rafforzare le metodologie tradizionali come l'analisi sensoriale o la verifica documentale, offrendo un livello di controllo fondato su evidenze chimiche misurabili (Cassago et al., 2021).

2.1.5 Rilevazione di Adulterazioni, Frodi e Modifiche non Dichiarate

Oltre a distinguere l'origine geografica, la metabolomica rappresenta uno strumento efficace per l'individuazione di adulterazioni e frodi alimentari. L'adulterazione di un prodotto, intesa come modifica deliberata e illecita della sua composizione naturale, lascia quasi sempre una "traccia" metabolica rilevabile attraverso le tecniche comunemente impiegate, in grado di individuare la presenza di metaboliti estranei o alterazioni nelle concentrazioni relative di quelli naturali (Mialon et al., 2023).
Selamat et al. (2021) hanno presentato una serie di risultati che dimostrano l'elevata sensibilità dell'approccio metabolomico untargeted nell'individuazione di falsificazioni, miscelazioni non dichiarate e adulterazioni volontarie in diversi comparti del settore agroalimentare. Nel caso della carne e dei prodotti ittici, l'analisi metabolica ha permesso di identificare biomarcatori capaci di distinguere tra specie diverse, di rilevare l'aggiunta fraudolenta di carni meno pregiate in preparazioni etichettate come "pure" e di monitorare alterazioni riconducibili a condizioni non conformi di conservazione o trasformazione. Anche nel comparto lattiero-caseario sono stati ottenuti risultati significativi: è stato possibile discriminare tra latte di diversa specie o

razza animale e rilevare la presenza occulta di latte vaccino in prodotti dichiarati come derivati di capra o bufala. Per quanto riguarda i prodotti trasformati, come la mozzarella di bufala e lo yogurt, l'impronta metabolica si è rivelata utile per risalire all'origine e distinguere campioni autentici da quelli non certificati o contraffatti. Nelle matrici vegetali, come frutta e ortaggi, la metabolomica ha mostrato un'analoga capacità discriminante. Le variazioni nei profili associate a biomarcatori discriminanti e indicativi dell'origine geografica e varietale, oltre che del metodo di coltivazione, si sono rivelate utili nell'individuazione di interventi di adulterazione e contraffazione. Analogamente, in prodotti come miele, caffè e tè, la metabolomica ha permesso di rilevare l'aggiunta non dichiarata di zuccheri, sciroppi o materie prime di minor pregio, supportando il controllo della veridicità delle indicazioni botaniche e geografiche riportate in etichetta (Selamat et al., 2021).

La metabolomica si dimostra utile anche per rilevare modifiche non dichiarate nei processi produttivi, che pur non configurandosi come vere e proprie frodi in senso giuridico, possono compromettere la qualità percepita del prodotto. Trattamenti come l'essiccazione o l'affumicatura aggiuntiva, l'impiego di mangimi specifici o l'irradiazione possono infatti alterare significativamente il profilo metabolico rispetto a quello atteso per un alimento standard (Utpott et al., 2022).

Inoltre, nel contesto della valutazione degli organismi geneticamente modificati (OGM), la metabolomica consente di confrontare varietà tradizionali e geneticamente modificate, evidenziando le differenze nei rispettivi profili metabolici (Benevenuto et al., 2022). Tali analisi risultano particolarmente rilevanti per garantire la conformità di filiere dichiarate esenti da OGM o da modificazioni non autorizzate.

In sintesi, ogni qualvolta un alimento subisce modifiche occulte nella composizione o nel processo produttivo, la metabolomica si conferma un mezzo affidabile per rilevarle, attraverso il confronto tra il profilo sospetto e un riferimento autentico. Studi consolidati confermano che questo approccio è in grado di identificare ingredienti adulteranti o componenti non dichiarati che minano l'autenticità e la qualità dei prodotti agroalimentari (Selamat et al., 2021). Di conseguenza, la metabolomica si configura come un potente deterrente contro le frodi e uno strumento essenziale per il controllo qualità lungo le catene di approvvigionamento.

2.1.6 Supporto alla Certificazione e Tracciabilità

Come descritto in precedenza, la definizione dei profili metabolici consente di distinguere l'origine geografica di un prodotto e di smascherare eventuali adulterazioni. Questa capacità rende la metabolomica un supporto scientifico di grande valore per i sistemi di certificazione

della qualità e tracciabilità alimentare. Certificazioni come DOP, IGP e DOC, fondate su disciplinari produttivi rigorosi e sull'origine garantita, possono beneficiare dell'introduzione di verifiche analitiche basate sul fingerprint metabolomico, da affiancare ai metodi tradizionali, quali controlli documentali, ispezioni visive e analisi chimiche mirate, al fine di confermare in modo oggettivo l'appartenenza di un campione alla categoria certificata.

Dal punto di vista istituzionale, l'impiego della metabolomica sta suscitando un interesse crescente anche in ambito normativo e giuridico. La possibilità di dimostrare scientificamente il legame tra un prodotto e il suo territorio, attraverso l'impronta metabolica, costituisce un'opportunità concreta per rafforzare la tutela legale delle denominazioni di origine, sia nella fase di registrazione, sia nei contenziosi contro le contraffazioni (Cassago et al., 2021). I dati metabolomici potrebbero essere impiegati dai consorzi di tutela per dimostrare l'unicità chimica del proprio prodotto rispetto a imitazioni, fornendo evidenze quantitative oggettive difficilmente contestabili. Inoltre, studi metabolomici condotti su prodotti candidati alla registrazione DOP o IGP possono mettere in luce, in modo solido, la correlazione tra qualità e territorio, contribuendo a rafforzare il dossier tecnico-scientifico necessario all'approvazione ufficiale (Cassago et al., 2021).

In ambito commerciale, la metabolomica può assumere anche una valenza strategica di marketing, in quanto comunicare al consumatore che l'autenticità di un alimento è stata verificata tramite tecnologie all'avanguardia rappresenta un elemento distintivo, in grado di accrescere la fiducia nel prodotto e valorizzare il brand territoriale. Non a caso, diverse analisi hanno mostrato che i prodotti certificati per origine raggiungono spesso prezzi e volumi di vendita superiori rispetto ai loro equivalenti non certificati, a conferma del fatto che l'investimento in qualità e autenticità è in grado di generare un ritorno economico tangibile (Cassago et al., 2021).

2.2 La spettroscopia NMR per la definizione dei profili metabolici dei Prodotti Agroalimentari

La Spettroscopia da Risonanza Magnetica Nucleare o NMR (*Nuclear Magnetic Resonance*) rappresenta una delle tecniche analitiche più avanzate e versatili per lo studio dei profili metabolici dei prodotti agroalimentari (Emwas et al., 2019, Sobolev et al., 2022).

Essa si basa sul comportamento magnetico di nuclei atomici specifici (principalmente H-1, C-13, P-31 e N-15), che, posti all'interno di un campo magnetico statico di elevata intensità, si orientano in direzioni definite rispetto al campo stesso. Quando questi nuclei vengono esposti a impulsi di radiofrequenza, assorbono energia specifica (frequenza di risonanza), passando da uno stato di minore energia a uno stato eccitato. Al termine dell'impulso, il ritorno dei nuclei al loro stato iniziale provoca un rilascio dell'energia assorbita sotto forma di segnali radio, definiti *Free Induction Decay* (FID). Questi segnali, rilevati e convertiti tramite Trasformata di Fourier, generano uno spettro NMR, il quale contiene informazioni dettagliate sulla composizione chimica e sulla struttura molecolare del campione analizzato. Ogni segnale presente nello spettro è caratterizzato da uno specifico spostamento chimico (*chemical shift*, δ), che fornisce indicazioni sull'ambiente molecolare in cui ciascun nucleo è inserito, e da un'intensità proporzionale alla concentrazione dei nuclei responsabili del segnale stesso (Jacobsen, 2007).

L'interpretazione combinata di questi parametri consente di ottenere informazioni dettagliate sia sulla natura chimica dei metaboliti presenti, sia sulla loro quantità relativa all'interno del campione analizzato. In particolare, lo shift chimico permette di identificare i gruppi funzionali e di dedurre la struttura molecolare delle sostanze, anche in matrici complesse. L'intensità del segnale, invece, consente una quantificazione relativa dei metaboliti e, qualora si utilizzino standard interni, anche una loro quantificazione assoluta (Picone, 2024). La ricostruzione dello spettro, ovvero del profilo metabolico completo, consente inoltre di ottenere una rappresentazione dell'impronta digitale chimica del campione — la cosiddetta *bio-fingerprint* — che riflette l'influenza congiunta di fattori genetici, ambientali, agronomici e tecnologici sulla sua composizione metabolica. Tale impronta risulta particolarmente utile per finalità di autenticazione, caratterizzazione e tracciabilità dei prodotti agroalimentari (Sobolev et al., 2022).

Grazie alla possibilità di applicare sia approcci mirati (*targeted*) sia non mirati (*untargeted*), l'NMR si rivela uno strumento estremamente flessibile per la determinazione di metaboliti chiave o per la valutazione globale della composizione del campione (Consonni & Cagliani,

2019). L'analisi delle *bio-fingerprints* permette di discriminare tra matrici provenienti da origini geografiche o filiere differenti, evidenziando eventuali non conformità o adulterazioni (Sobolev et al., 2021; García-Pérez et al., 2024).

In questo contesto, la spettroscopia NMR si configura come una tecnica strategica per la valorizzazione e la tutela delle produzioni di qualità, in quanto consente di ricostruire l'origine, i processi produttivi e la conformità del prodotto rispetto agli standard attesi, offrendo una *firma chimica globale* utile lungo l'intera filiera.

2.2.1 Vantaggi e Limiti della Spettroscopia NMR in metabolomica

Rispetto ad altre tecniche analitiche comunemente impiegate nella metabolomica alimentare, quali, ad esempio la Spettrometria di Massa, la Spettroscopia NMR presenta alcuni vantaggi significativi.

Tra i principali punti di forza di questa tecnica vi sono la sua natura intrinsecamente non distruttiva e la possibilità di effettuare analisi senza la necessità di pretrattamenti (come la ionizzazione o la derivatizzazione chimica), che potrebbero alterare la composizione chimica o la struttura molecolare della matrice esaminata (Jacobsen, 2007; Emwas et al., 2019). Questo significa che un campione può essere riutilizzato per più misurazioni successive, anche in condizioni sperimentali differenti o sfruttando tecniche analitiche complementari. Inoltre, il fatto che il profilo metabolico osservato rifletta fedelmente la composizione originale del prodotto, senza interferenze introdotte dal processo analitico, rappresenta un elemento di grande valore nella caratterizzazione di alimenti complessi e nelle analisi a supporto dell'autenticazione e della tracciabilità (Sobolev et al., 2022).

Un aspetto particolarmente rilevante dell'NMR è la sua versatilità, che permette di analizzare campioni sia allo stato liquido che solido, variando la sonda nel quale il campione viene inserito.

Un altro punto di forza della Spettroscopia NMR è rappresentato dalla sua elevata riproducibilità e precisione, attribuibile alla minore influenza delle variabili analitiche e delle procedure preparative. Diversi studi hanno confermato che i risultati ottenuti tramite NMR presentano una bassa variabilità interlaboratorio, a dimostrazione della robustezza e affidabilità del metodo (Gallo et al., 2020). Ciò rende questa tecnica particolarmente adatta per applicazioni ufficiali nel controllo di qualità in ambito alimentare. Infine, la Spettroscopia NMR offre la possibilità di effettuare analisi rapide, con tempi di preparazione del campione ridotti o addirittura assenti, il che costituisce un ulteriore vantaggio. Questo aspetto contribuisce a un significativo risparmio di tempo e risorse, risultando particolarmente vantaggioso nei contesti ad alta produttività o negli screening metabolomici su larga scala (Emwas et al., 2019).

Tuttavia, la Spettroscopia NMR presenta anche alcuni limiti, che devono essere attentamente considerati nella scelta della tecnica analitica più appropriata. La principale limitazione riguarda la sensibilità relativamente bassa, che risulta decisamente inferiore rispetto a quella della Spettrometria di Massa. Quest'ultima, infatti, è in grado di rilevare e quantificare metaboliti presenti a concentrazioni molto più basse — anche da 10 a 100 volte inferiori — rendendola preferibile per analisi mirate su composti in tracce o poco abbondanti (Emwas et al., 2019).

Un ulteriore limite della Spettroscopia NMR è rappresentato dagli elevati costi iniziali, legati all'acquisto e alla manutenzione degli strumenti, in particolare degli spettrometri ad alto campo magnetico. Inoltre, la tecnica richiede la presenza di personale altamente specializzato per l'analisi e l'interpretazione dei dati, un fattore che può limitarne l'impiego nei laboratori meno attrezzati o la diffusione in ambito industriale e nei sistemi di controllo alimentare (Sobolev et al., 2022).

2.2.2 Principali tecniche in Spettroscopia NMR

La Spettroscopia NMR offre un'ampia gamma di possibilità applicative grazie alla disponibilità di diverse modalità sperimentali, ciascuna delle quali si adatta a specifiche esigenze analitiche nel contesto agroalimentare (Emwas et al., 2019).

2.2.2.1 Spettroscopia ^{1}H NMR

La spettroscopia ^{1}H-NMR rappresenta la tecnica più comunemente utilizzata negli studi metabolomici, grazie all'elevata abbondanza isotopica naturale (~99%) degli atomi di idrogeno nei composti organici e alle loro favorevoli caratteristiche fisiche.

L'applicazione standard di questa tecnica prevede l'analisi di campioni in fase liquida, condizione che consente l'ottenimento di spettri ad alta risoluzione grazie alla mobilità molecolare e all'omogeneità del mezzo. Nel contesto alimentare, ciò comporta la solubilizzazione o l'estrazione della matrice — ad esempio succhi, bevande, oli, estratti vegetali, farine o formaggi — generalmente tramite tamponi in acqua deuterata (D_2O) o miscele idro-organiche, al fine di isolare i metaboliti e renderli analizzabili (Picone, 2024).

La spettroscopia ^{1}H-NMR è altamente automatizzabile, affidabile e rapida. I tempi di acquisizione possono essere molto brevi (anche pochi minuti) e, grazie all'automazione degli strumenti moderni, è possibile analizzare in sequenza numerosi campioni, rendendo questa metodica particolarmente adatta a studi su larga scala.

Poiché i segnali registrati riflettono direttamente la distribuzione e la concentrazione dei protoni nelle molecole, è possibile ottenere una quantificazione accurata dei metaboliti, a condizione di garantire condizioni sperimentali stabili, l'applicazione di protocolli standardizzati e una corretta ottimizzazione dei parametri di acquisizione. Un aspetto tecnico cruciale riguarda la soppressione del segnale dell'acqua, data la predominanza del solvente acquoso nei campioni biologici. Sebbene esistano diverse strategie per attenuarlo, molte di esse richiedono competenze avanzate e una calibrazione precisa dei parametri, rendendone l'applicazione complessa per operatori meno esperti.

Tra le principali limitazioni della spettroscopia ^{1}H-NMR si segnala la ristretta finestra di *chemical shift*, che può comportare sovrapposizioni tra i segnali (*peak overlap*), rendendo più difficile l'identificazione e la quantificazione dei metaboliti. Una strategia efficace per mitigare questo problema è l'impiego di spettrometri ad alto campo magnetico, che migliorano la risoluzione spettrale in modo proporzionale al campo applicato. Tuttavia, l'elevato costo di tali strumenti ne limita la diffusione, e il campo a 600 MHz viene spesso considerato il miglior compromesso tra prestazioni e accessibilità. Un'ulteriore soluzione consiste nell'utilizzo di tecniche di eccitazione selettiva e di esperimenti NMR bidimensionali (come COSY, TOCSY, DOSY, HSQC, HMBC), che generano mappe di correlazione su due assi, facilitando la distinzione tra segnali sovrapposti e la ricostruzione delle relazioni strutturali tra i nuclei. Questo approccio consente un'attribuzione più precisa dei segnali spettrali a specifici metaboliti.

Nonostante il potenziale informativo offerto dalla NMR multidimensionale, per ragioni economiche e operative molti studi metabolomici si limitano ancora all'utilizzo di esperimenti monodimensionali (Emwas et al., 2019).

2.2.2.2 Spettroscopia ^{13}C NMR

Rispetto alla spettroscopia ^{1}H-NMR, la spettroscopia ^{13}C-NMR offre una maggiore dispersione dello shift chimico (fino a circa 200 ppm), garantendo un vantaggio significativo in termini di risoluzione spettrale. Tuttavia, la bassa abbondanza naturale del nucleo ^{13}C (circa 1,1%) e la minore sensibilità del segnale ne limitano l'impiego diretto nella metabolomica (Emwas et al., 2019).

La ^{13}C-NMR risulta utile per la caratterizzazione di componenti come polisaccaridi, fibre e lipidi. L'impiego di esperimenti multidimensionali, come HSQC e HMBC, consente di superare le limitazioni legate alla bassa sensibilità, acquisendo la risonanza del carbonio attraverso il canale del protone. Questa strategia permette di combinare le informazioni fornite dal ^{13}C con

quelle del ^{1}H, migliorando significativamente l'identificazione di metaboliti in matrici complesse (Truzzi et al., 2021; Wu et al., 2022).

2.2.2.3 La Spettroscopia NMR bidimensionale (2D)

La spettroscopia NMR bidimensionale (2D) rappresenta un'evoluzione significativa rispetto alla classica NMR monodimensionale (1D), offrendo maggiori possibilità di identificazione molecolare e determinazione strutturale. Sebbene richieda tempi di acquisizione più lunghi e competenze interpretative avanzate, la NMR 2D consente una migliore separazione dei segnali e una più precisa identificazione dei metaboliti, soprattutto in matrici complesse. Il suo potenziale è ampiamente riconosciuto in ambito metabolomico, dove rappresenta uno strumento fondamentale per affrontare il problema della sovrapposizione spettrale tipico degli spettri 1D, grazie alla distribuzione dei segnali lungo una seconda dimensione basata su proprietà fisiche ortogonali, come la costante di accoppiamento o il tipo di legame.
Tra gli esperimenti più utilizzati in metabolomica vi sono i 2D $^{1}H–^{1}H$ come TOCSY, COSY e NOESY, oltre ai classici esperimenti eteronucleari come $^{1}H–^{13}C$ HSQC e HMBC, fondamentali per l'analisi delle connessioni tra atomi di idrogeno e carbonio. Altre tecniche 2D impiegate includono la DOSY (diffusion ordered spectroscopy), utile per separare i metaboliti in base al loro coefficiente di diffusione, e la J-resolved NMR, che aiuta a distinguere i segnali sovrapposti attraverso la separazione delle costanti di accoppiamento (Emwas et al., 2019, Moco, 2022).

2.2.2.4 La Spettroscopia NMR con rotazione ad "Angolo Magico" (Magic Angle Spinning NMR)

Tra le innovazioni più significative introdotte nel campo della spettroscopia NMR applicata alla metabolomica figura la High-Resolution Magic-Angle Spinning (HRMAS), una tecnica progettata per l'analisi di campioni semisolidi o solidi idratati, tipicamente non accessibili mediante le metodologie NMR in soluzione. Grazie alla rotazione del campione ad alta velocità a un angolo di 54,74° rispetto al campo magnetico statico (il cosiddetto *angolo magico*), l'HRMAS consente di minimizzare gli effetti delle interazioni dipolari e dell'anisotropia dello spostamento chimico (Chemical-Shift-Anisotropy - CSA), che sono tra i principali responsabili dell'allargamento delle linee spettrali nei materiali solidi. Questo accorgimento rende possibile l'acquisizione di spettri ad alta risoluzione anche da matrici complesse ed eterogenee, mantenendo l'integrità fisica e chimica del campione (Mazzei & Piccolo, 2017; Emwas et al., 2019). Un principio analogo è alla base della tecnica Cross Polarization Magic Angle Spinning (CP-MAS), specificamente applicata alla spettroscopia ^{13}C-NMR su campioni solidi. Questa

metodologia consente il trasferimento di magnetizzazione dai nuclei ^{1}H ai ^{13}C, incrementando l'intensità del segnale del carbonio. Grazie alla rotazione all'angolo magico (MAS), è possibile ridurre significativamente l'effetto dell'anisotropia chimica e degli accoppiamenti dipolari, che in campioni statici causerebbero un marcato allargamento delle linee spettrali. In questo modo si ottengono spettri più risoluti, sebbene con una risoluzione inferiore rispetto a quella degli spettri in fase liquida, permettendo comunque l'analisi strutturale di materiali solidi amorfi o complessi, inclusi tessuti vegetali, matrici lignocellulosiche o fibre alimentari (Conte et al., 2004).

La spettroscopia 13C CP-MAS NMR si è dimostrata particolarmente utile nella caratterizzazione strutturale dei cippati di biomassa vegetale. Tali cippati, ottenuti dalla triturazione di residui legnosi, presentano una granulometria variabile e un'origine eterogenea, che determinano le loro caratteristiche fisico-chimiche e influenzano le modalità di impiego (Bonanomi et al., 2013). Dal punto di vista analitico, la composizione dei cippati, dominata da cellulosa, emicellulose e lignina, rappresenta un modello ideale per l'applicazione di tecniche NMR allo stato solido. In particolare, la spettroscopia 13C CP-MAS consente l'identificazione delle componenti principali senza necessità di pretrattamenti estrattivi, permettendo di distinguere tra le forme cristallina e amorfa della cellulosa, di individuare i gruppi metossilici e aromatici della lignina e di rilevare i segnali specifici delle emicellulose (Cipriano et al., 2020; Conte et al., 2004; Kostryukov et al., 2021). Inoltre, l'analisi 13C CP-MAS permette di stimare in modo semi-quantitativo la composizione carboniosa complessiva del campione, fornendo indicazioni preziose sulla distribuzione dei principali gruppi funzionali.

2.2.3 La Spettroscopia HR MAS NMR in metabolomica alimentare

Tradizionalmente, l'analisi NMR dei campioni alimentari solidi presenta intrinseche difficoltà legate alla perdita di risoluzione spettrale, dovuta all'allargamento delle linee di segnale. Come discusso in precedenza, questo fenomeno è causato da interazioni anisotrope, quali accoppiamenti dipolari (sia omonucleari che eteronucleari), effetti quadrupolari, anisotropia dello spostamento chimico (CSA) e variazioni locali della suscettibilità magnetica. Nonostante l'introduzione di strategie correttive, queste interazioni continuano a ostacolare l'acquisizione di spettri ad alta risoluzione. (Mazzei & Piccolo, 2017).

Per superare tali limiti, una strategia consolidata prevede la dissoluzione del campione in un solvente, trasformandolo in una matrice liquida. In soluzione, infatti, l'isotropia del moto molecolare consente di mediare naturalmente le interazioni anisotrope, permettendo di ottenere spettri NMR ad alta risoluzione. Tuttavia, il trasferimento allo stato liquido comporta operazioni

di estrazione, purificazione e concentrazione, che non solo sono laboriose e dispendiose in termini di tempo, ma possono anche introdurre artefatti analitici dovuti alla perdita o degradazione di metaboliti labili, alterando la rappresentatività del profilo metabolico originario (Mazzei & Piccolo, 2017). La tecnica High-Resolution Magic Angle Spinning (HR-MAS), introdotta alla fine degli anni ’90, nasce proprio con l’obiettivo di combinare i vantaggi dell’NMR allo stato liquido e solido, superandone i rispettivi limiti. Nei protocolli applicativi, il campione viene ridotto in piccoli frammenti e inserito in un rotore sigillato (solitamente in zirconia, con diametro ≤ 4 mm), contenente una minima quantità di solvente deuterato (es. D_2O), utile a garantire l’idratazione e la stabilità del campo magnetico (lock). Il rotore viene quindi fatto ruotare a velocità moderata (3–6 kHz) attorno a un asse inclinato di 54,7° rispetto al campo magnetico statico. Questa configurazione consente di minimizzare le interazioni anisotrope, riducendo i couplings dipolari H–H da decine di kHz (tipici dei solidi rigidi) a poche centinaia di Hz. Si ottengono così spettri ad alta risoluzione anche da matrici eterogenee e intatte, con qualità comparabile a quella degli esperimenti in soluzione (Mazzei & Piccolo, 2017). La qualità degli spettri ottenuti tramite HR-MAS risulta, nella maggior parte dei casi, comparabile a quella dei corrispondenti estratti liquidi, con valori di larghezza a metà altezza (FWHM) sovrapponibili e una completa conservazione delle molteplicità dei segnali (Valentini et al., 2011). La risoluzione spettrale raggiunta è tale da consentire l'esecuzione della maggior parte degli esperimenti NMR monodimensionali e bidimensionali, sia omonucleari che eteronucleari, anche su campioni intatti o minimamente trattati. La tecnica si presta particolarmente all'analisi di sistemi molecolari complessi, come polimeri, proteine di membrana e tessuti biologici, e può essere ulteriormente potenziata attraverso l'impiego di sequenze di impulso modificate, che permettono di distinguere tra componenti mobili e meno mobili all'interno di matrici idratate. Sebbene le prime applicazioni della HR-MAS si siano concentrate principalmente nel campo biomedico e dei materiali, numerosi studi hanno evidenziato il suo potenziale anche nell’ambito della chimica agraria e della caratterizzazione di matrici agroalimentari (Santos et al., 2015).

Le matrici solide e semisolide — caratterizzate dalla coesistenza di componenti fibrose o granulari (come pareti cellulari, amido, lignina) e di una fase liquida intracellulare — sono quelle che maggiormente beneficiano dell’analisi HR-MAS. In questi casi, la tecnica consente di sondare simultaneamente entrambe le fasi, fornendo un profilo metabolico più completo rispetto alla NMR tradizionale, che richiederebbe di separare le componenti o di procedere a filtrazione/estrazione selettiva (Santos et al., 2015). In definitiva, nel contesto della metabolomica alimentare, la HR-MAS rappresenta uno strumento potente e complementare alle

tecniche classiche, permettendo lo studio degli alimenti in condizioni prossime a quelle reali, senza introdurre bias analitici dovuti al trattamento del campione.

2.2.3.1 Applicazioni della spettroscopia HR-MAS NMR in metabolomica alimentare

Negli ultimi anni, la spettroscopia HR-MAS NMR (High Resolution Magic Angle Spinning Nuclear Magnetic Resonance) si è affermata come una tecnica analitica estremamente efficace per la caratterizzazione non distruttiva di prodotti agroalimentari. Grazie alla possibilità di analizzare campioni semi-solidi o solidi idratati senza estrazione, la tecnica consente di ottenere un'impronta metabolica dettagliata direttamente sul tessuto integro, preservando la matrice nella sua forma naturale. Le sue applicazioni si estendono oggi a numerose categorie alimentari, dalla frutta e verdura ai cereali, dai latticini fino ai prodotti ittici, alla carne e agli oli, dimostrandosi particolarmente utile nei contesti di autenticazione, tracciabilità geografica e controllo qualità. (Santos et al., 2015; Corsaro et al.,2016; Mazzei & Piccolo, 2017; Ragone et al., 2020).

Frutta e ortaggi

L'HR-MAS NMR ha mostrato eccellenti capacità nel monitoraggio dei cambiamenti metabolici associati alla maturazione e alla lavorazione post-raccolta. Nei primi studi sul mango (Mangifera indica L.), è stato possibile tracciare con precisione l'evoluzione del profilo chimico della polpa lungo il processo di maturazione, osservando una marcata riduzione dell'acido citrico e l'accumulo di GABA, niacina, alanina e zuccheri semplici. Analogamente, nel pomodoro (Solanum lycopersicum), l'analisi ha permesso la discriminazione tra diverse componenti tissutali (polpa, buccia, semi, purea) e lo studio delle variazioni metaboliche associate a cultivar e stadio fenologico.

La tecnica è stata inoltre impiegata con successo per distinguere campioni ortofrutticoli secondo l'origine geografica e varietale. L'analisi multivariata di spettri HR-MAS ha evidenziato, ad esempio, differenze significative nei profili metabolici dei pomodorini ciliegino di Pachino IGP rispetto a campioni di origine incerta, così come nei peperoni dolci italiani (Capsicum annuum L.), distinguendo le varietà "Corno" e "Cuneo" coltivate in diverse regioni. Risultati analoghi sono stati ottenuti per i limoni Interdonato IGP, nei quali l'HR-MAS ha permesso la separazione chiara tra campioni siciliani certificati e varietà turche, evidenziando differenze nei contenuti di zuccheri, acidi organici e amminoacidi.

Cereali e derivati

Nel settore cerealicolo, la spettroscopia HR-MAS si è rivelata uno strumento sensibile per l'analisi di farine, impasti e prodotti trasformati. Già nel 1998, l'analisi di spettri 1H HR-MAS

ha permesso di discriminare farine di grano duro provenienti da diverse aree del Sud Italia, basandosi su variazioni nei contenuti di lipidi e polisaccaridi. Studi successivi hanno esplorato le fasi di panificazione, rilevando metaboliti come acido lattico, succinico e acetico nelle briciole di pane, assenti nella farina di partenza, e monitorando l'idrolisi enzimatica dei carboidrati in tempo reale.

Esperimenti più recenti hanno approfondito la mobilità dell'acqua nei granuli di amido (frumento, mais, patata), mentre l'analisi dei semi di orzo ha evidenziato il ruolo del contenuto in β-glucani nel determinare il profilo metabolico durante la maturazione. L'HR-MAS è stata anche utilizzata per valutare la risposta a stress ambientali nel riso e per studiare l'equivalenza compositiva tra colture convenzionali e OGM, come nel caso di fagioli transgenici resistenti al virus BGMV.

Oli e spezie

L'applicazione della HR-MAS NMR all'olio extravergine di oliva ha fornito risultati significativi nella determinazione del profilo lipidico e nella distinzione tra oli DOP siciliani (es. Val di Mazara, Valli Trapanesi, Valdemone). Utilizzando protocolli rapidi basati su relazioni spettrali tra gruppi metilici e metilenici degli acidi grassi e i protoni del glicerolo, è stato possibile stimare con precisione le percentuali di acido oleico, linoleico, linolenico e acidi grassi saturi.

Nel caso delle spezie, la tecnica ha consentito la caratterizzazione dettagliata del peperone dolce, rivelando composti minori come trigonellina, colina e derivati piridinici. L'analisi PLS-DA ha confermato la capacità di discriminare le varietà e l'origine geografica, evidenziando il contributo di zuccheri, acidi grassi e acidi organici nella classificazione.

Latticini e formaggi

Nel settore lattiero-caseario, la HR-MAS si è distinta per la sua capacità di studiare i prodotti nel loro stato nativo. L'applicazione al Parmigiano Reggiano ha permesso di monitorare l'evoluzione metabolica durante la stagionatura, con accumulo progressivo di amminoacidi liberi e acidi organici. In studi successivi, la tecnica ha permesso la distinzione tra formaggi Emmental di diversa provenienza geografica e la caratterizzazione metabolica della Mozzarella di Bufala Campana, utilizzando sequenze di impulso modificate e analisi statistica multivariata. Marker come β-lattosio, β-galattosio, acido acetico e isobutanolo si sono rivelati utili per distinguere i campioni freschi da quelli stagionati.

Prodotti ittici e carni

Infine, l'HR-MAS NMR ha trovato ampio impiego nell'analisi di pesce e carne. Sui filetti di salmone e salmerino, è stato possibile quantificare DHA, EPA e altri acidi grassi ω-3, insieme

a metaboliti come anserina, taurina e creatina. Le sequenze diffusion-edited e CPMG hanno consentito di distinguere tra composti a basso e alto peso molecolare. Nel salmone affumicato, sono stati identificati oltre 160 metaboliti; in più, la tecnica si è dimostrata efficace nel monitorare l'effetto della β-irradiamento sulla conservazione del prodotto.

Per quanto riguarda le carni, l'analisi HR-MAS ha permesso di discriminare la provenienza geografica di carne ovina e bovina, e di identificare marcatori associati alla razza o al taglio muscolare (es. longissimus dorsi vs semitendinosus). Un aspetto critico è risultato essere la regolazione del pH nei campioni, fondamentale per garantire la comparabilità spettrale e l'accuratezza delle analisi statistiche.

2.3 Metodi chemometrici classici per l'interpretazione degli Spettri NMR

La spettroscopia NMR rappresenta una tecnica analitica particolarmente efficace nell'ambito della metabolomica *untargeted*, finalizzata alla ricostruzione del *fingerprint* metabolico globale di una matrice alimentare (Ellis et al., 2012; Ragone et al., 2020; Musio et al., 2022; Sobolev et al., 2022). Tale impronta chimica, generata dallo spettro NMR, costituisce una rappresentazione complessa e informativamente densa della composizione molecolare del campione.

Le procedure convenzionali di preprocessing permettono di convertire tale informazione spettrale in una matrice numerica ad alta dimensionalità, tipicamente una *bucket table* – ovvero una tabella che riporta l'integrazione dell'intensità dei segnali in intervalli regolari dello spettro (buckets), consentendo così l'elaborazione statistica del dato (Heyman et al., 2012; Antonicelli et al., 2021; Sobolev et al., 2022).

Per interpretare in modo efficace questi dataset complessi, è necessario ricorrere alla chemometria, intesa come l'applicazione di strumenti matematici e statistici per l'analisi di dati chimici, con l'obiettivo di estrarre conoscenza da sistemi altamente informativi e multidimensionali (Aleixandre-Tudo et al., 2022).

In ambito metabolomico, l'impiego della chemometria si rivela essenziale per affrontare la variabilità intrinseca delle matrici alimentari, ridurre la dimensionalità dei dati e mettere in evidenza pattern sistematici o relazioni latenti tra campioni e variabili (Gowda & Raftery, 2021). I metodi chemiometrici si articolano principalmente in due categorie: gli approcci esplorativi, impiegati per individuare strutture intrinseche nei dati in assenza di informazioni a priori, e gli approcci predittivi, volti a costruire modelli statistici per classificare nuovi campioni o stimarne proprietà sulla base di etichette note (Andre et al., 2020; Sobolev et al., 2022).

Tali strumenti consentono, ad esempio, di discriminare i campioni in base all'origine botanica o geografica, al metodo di produzione, ai processi di trasformazione e alle condizioni di conservazione (Mazzei & Piccolo, 2017; Musio et al., 2022). In questo contesto, la spettroscopia NMR, integrata con metodi chemiometrici, si configura come uno strumento strategico per applicazioni orientate all'autenticazione, alla tracciabilità e alla certificazione della qualità (Musio et al., 2022, Sobolev et al., 2022).

2.3.1 Analisi Esplorativa dei Dati: PCA e HCA

Le principali tecniche esplorative impiegate in ambito chemiometrico sono la *Principal Component Analysis* (PCA) e la *Hierarchical Cluster Analysis* (HCA), entrambe ampiamente utilizzate per l'analisi preliminare di dataset complessi, come i profili metabolici generati mediante spettroscopia NMR (Mazzei et al., 2010; Gowda & Raftery, 2021; Sobolev et al., 2022).

La *Principal Component Analysis* (PCA) è una tecnica multivariata non supervisionata che ha come obiettivo principale la riduzione della dimensionalità dei dati, conservando quanto più possibile l'informazione originaria e facilitando la visualizzazione e l'interpretazione delle strutture latenti.

Questo risultato si ottiene mediante una trasformazione ortogonale delle variabili originarie in un nuovo insieme di variabili non correlate, dette componenti principali (Principal Components, PC). Queste sono anche definite componenti latenti, poiché non corrispondono a variabili osservabili direttamente, ma rappresentano combinazioni lineari delle variabili iniziali. In altri termini, si tratta di nuove variabili "astratte" che sintetizzano le principali tendenze presenti nel dataset, contribuendo a rivelarne le strutture sottostanti.

Dal punto di vista matematico, la PCA risolve un problema di autovalori e autovettori sulla matrice di covarianza (o, in alternativa, sulla matrice di correlazione) dei dati: gli autovettori identificano le direzioni lungo cui la varianza è massima, mentre gli autovalori quantificano la quota di varianza spiegata da ciascuna direzione. La prima componente principale (PC1) è orientata lungo la direzione che cattura la massima variabilità, la seconda (PC2) lungo una direzione ortogonale alla prima, e così via (Worley & Powers, 2013).

Il risultato è una rappresentazione compatta e informativa dei dati in uno spazio a dimensionalità ridotta, utile per esplorare pattern, differenze tra gruppi e individuare outlier. Dal punto di vista grafico, i risultati della PCA vengono generalmente rappresentati attraverso due diagrammi fondamentali: lo score plot, che visualizza la distribuzione dei campioni nello spazio delle componenti principali e il loading plot, che mostra il contributo di ciascuna variabile originaria – ad esempio i bucket spettrali NMR – a ciascuna componente (Corsaro et al., 2022).

In ambito metabolomico, la PCA si rivela particolarmente efficace nell'individuare cluster naturali di campioni – come alimenti con profili metabolici simili – e nell'evidenziare le variabili spettrali più rilevanti per la separazione tra gruppi. Queste ultime possono essere individuate mediante il loading plot e il calcolo del contributo percentuale di ciascuna variabile

alla costruzione delle componenti principali, e successivamente associate a metaboliti discriminanti attraverso tecniche di annotazione spettrale e confronto con database metabolomici (Worley & Powers, 2013; Corsaro et al., 2022).

Va tuttavia osservato che, poiché la PCA massimizza esclusivamente la varianza totale del dataset, la separazione tra gruppi può essere influenzata da fonti di varianza dominanti che non riflettono necessariamente differenze biologiche o metaboliche, ma possono essere legate a rumore, effetti tecnici o variabili confondenti. Per questo motivo, pur rappresentando uno strumento esplorativo essenziale, la PCA deve essere interpretata con cautela e, ove necessario, affiancata da metodi supervisionati – come la PLS-DA o l'OPLS-DA – per ottimizzare la discriminazione tra le classi di interesse.

Infine, per garantire l'affidabilità statistica dell'analisi, è fondamentale disporre di un numero adeguato di osservazioni rispetto al numero di variabili. In situazioni di sovradimensionamento – in cui il numero di variabili supera di gran lunga quello dei campioni – aumenta il rischio che le componenti principali siano influenzate dal rumore, compromettendo la validità dei risultati (Worley & Powers, 2013).

La *Hierarchical Cluster Analysis* (HCA) è una tecnica multivariata non supervisionata frequentemente utilizzata nei contesti di metabolomica e chemometria per esplorare la struttura intrinseca dei dati, in particolare quando si desidera individuare raggruppamenti naturali di campioni in base alla loro similarità globale (Mazzei et al., 2010; Sobolev et al., 2022; Gowda & Raftery, 2021; Corsaro et al., 2022).

A differenza della PCA, che si basa sull'analisi della varianza, la HCA utilizza misure di distanza o dissimilarità tra osservazioni per costruire una struttura gerarchica ad albero, chiamata dendrogramma. Il processo inizia con il calcolo di una matrice di distanza (di norma sulla base della distanza euclidea), seguita da un algoritmo di aggregazione che unisce progressivamente i campioni più simili in cluster. Questo procedimento prosegue fino a includere tutti i campioni in un unico cluster complessivo. Il dendrogramma risultante offre una rappresentazione visiva del livello di somiglianza tra i campioni: i campioni molto simili si uniscono nei rami inferiori dell'albero, mentre quelli più dissimili si aggregano solo a livelli superiori. Questa visualizzazione è particolarmente utile per mettere in evidenza pattern latenti, verificare la coerenza di classificazioni a priori, rilevare outlier o sottogruppi significativi (Worley & Powers, 2013).

Nel contesto della spettroscopia NMR applicata al settore agroalimentare, la HCA rappresenta uno strumento efficace per discriminare prodotti in base al loro profilo metabolico, ad esempio distinguendo alimenti provenienti da diverse regioni geografiche o processi (Mazzei et al.,

2010; Corsaro et al., 2022). Pur non fornendo una riduzione di dimensionalità come la PCA, la HCA costituisce un approccio complementare e potente per esplorare la struttura dei dati, utile per generare ipotesi preliminari e guidare successive analisi supervisionate.

2.3.2 Analisi Predittiva: PLS-DA e OPLS-DA

Dopo la fase esplorativa, in cui si ricercano pattern latenti e raggruppamenti naturali nei dati, l'analisi chemiometrica può essere ulteriormente approfondita attraverso approcci supervisionati, finalizzati a costruire modelli predittivi in grado di distinguere tra classi predefinite di campioni. In questo contesto, la *Partial Least Squares Discriminant Analysis* (PLS-DA) e la sua variante *Orthogonal PLS-DA* (OPLS-DA) rappresentano due delle tecniche più utilizzate in ambito metabolomico (Sobolev et al., 2022).

La *Partial Least Squares Discriminant Analysis* (PLS-DA) è una tecnica multivariata supervisionata che costruisce un modello predittivo capace di descrivere la relazione tra un insieme di variabili indipendenti (X), come i bucket spettrali NMR, e una variabile dipendente (Y) rappresentata da classi categoriali, ad esempio la provenienza geografica, la varietà botanica o il metodo di trasformazione del prodotto.

L'obiettivo principale della PLS-DA è identificare componenti latenti che non solo spiegano la varianza nei dati X, ma che siano anche fortemente correlate con Y, cioè con le categorie da prevedere (Wold et al., 2001). La denominazione "minimi quadrati" deriva dal fatto che la tecnica cerca di costruire queste componenti latenti minimizzando la somma dei quadrati delle differenze tra i valori osservati e quelli predetti di Y, analogamente a quanto avviene nella regressione lineare. Tuttavia, a differenza dei metodi di regressione tradizionali, la PLS-DA riesce a trattare anche dataset altamente collineari e con più variabili che osservazioni, come avviene frequentemente nei dati metabolomici. Per queste sue caratteristiche, la PLS-DA è spesso descritta come una sorta di "PCA supervisionata", poiché combina la riduzione della dimensionalità con la capacità di massimizzare la separazione tra classi conosciute. Ciò la rende particolarmente adatta per l'analisi di dati complessi e rumorosi come quelli derivanti da spettroscopia NMR (Sobolev et al., 2022).

Oltre alla classificazione, un ulteriore punto di forza della PLS-DA è la possibilità di identificare variabili discriminanti mediante il calcolo dei *VIP scores* (*Variable Importance in Projection*). I VIP rappresentano una stima dell'importanza di ciascuna variabile nella costruzione del modello predittivo e nella separazione tra le classi (Wold et al., 2001), e sono frequentemente utilizzati per individuare i potenziali metaboliti rilevanti per la caratterizzazione del campione (Rivera-Pérez et al., 2023).

Essendo una tecnica supervisionata, la PLS-DA richiede una validazione rigorosa per evitare fenomeni di overfitting, cioè un'eccessiva adattabilità del modello ai dati di addestramento che ne comprometterebbe la generalizzazione. Tra le strategie di validazione più comuni si annoverano la *cross-validazione interna*, utile a determinare il numero ottimale di componenti latenti (Westerhuis et al., 2008), e il *test di permutazione* (*label shuffling*), che consente di verificare se la separazione osservata tra le classi sia statisticamente significativa oppure attribuibile al caso (Szymańska et al., 2012). Un modello ben validato dovrebbe restituire alti valori di varianza spiegata e capacità predittiva in cross-validazione, accompagnati da un p-value inferiore a 0.05 nel test di permutazione, a conferma della sua robustezza (Worley & Powers, 2013).

L'*Orthogonal Partial Least Squares-Discriminant Analysis* (OPLS-DA) rappresenta un'evoluzione avanzata della PLS-DA, concepita per migliorare l'interpretabilità e l'efficienza dei modelli di classificazione multivariata. Si tratta di un metodo supervisionato che introduce una scomposizione più strutturata della matrice dei predittori X, separando esplicitamente la componente di variabilità correlata alla variabile di risposta Y da quella ortogonale, ovvero indipendente da essa.

Dal punto di vista operativo, l'OPLS-DA proietta i dati in uno spazio latente suddiviso in componenti predittive, che descrivono la parte della varianza in X correlata a Y, e componenti ortogonali, che modellano la varianza sistematica in X non correlata a Y. Questa separazione consente di isolare e modellare in maniera più efficace la cosiddetta "varianza di disturbo", spesso dovuta a fattori confondenti o rumore sperimentale, che altrimenti rischierebbero di mascherare la reale separazione tra classi (Trygg & Wold, 2002).

Il principale vantaggio dell'OPLS-DA risiede nel fatto che la discriminazione tra i gruppi è concentrata nella prima componente predittiva, mentre le componenti ortogonali assorbono la variabilità interna alle classi, legata a fattori secondari. Questo approccio si traduce in score plot più chiari e interpretabili, con separazioni nette tra i gruppi. Inoltre, l'eliminazione della varianza non rilevante consente una stima più robusta dei VIP scores e dei coefficienti di regressione, facilitando l'identificazione di variabili discriminanti (Wold et al., 2001).

Dal punto di vista computazionale, l'OPLS-DA è un modello più parsimonioso, poiché richiede un numero minore di componenti latenti rispetto alla PLS-DA, ed è potenzialmente meno soggetto a overfitting, a condizione che venga validato correttamente mediante tecniche come la cross-validazione e i test di permutazione (Westerhuis et al., 2008).

2.3.3 Analisi univariata: confronto tra gruppi mediante ANOVA e test di Tukey

In ambito metabolomico, l'analisi univariata rappresenta un complemento utile alle tecniche multivariate, in quanto consente di valutare separatamente il contributo statistico di ciascuna variabile. Tra i test univariati più impiegati figura l'Analisi della Varianza (ANOVA, Analysis of Variance), che permette di confrontare le medie di una variabile tra due o più gruppi e di verificare se le differenze osservate siano statisticamente significative. In questo modo, è possibile stabilire se una determinata variabile – ad esempio un bucket spettrale NMR – contribuisca in modo rilevante alla discriminazione tra i gruppi (Mazzei et al., 2010; Saccenti et al., 2014).

A differenza dei metodi multivariati, come PCA o PLS-DA, che analizzano simultaneamente tutte le variabili tenendo conto delle loro correlazioni, l'approccio univariato isola l'effetto specifico di ogni variabile, riducendo il rischio che segnali rilevanti vengano mascherati da rumore o variabili non informative (Broadhurst & Kell, 2006; Saccenti et al., 2014). Tuttavia, l'analisi univariata presenta anche delle limitazioni: in particolare, richiede l'applicazione di correzioni per test multipli – come la False Discovery Rate (FDR) o la correzione di Bonferroni – per contenere il rischio di falsi positivi, dovuto all'elevato numero di confronti simultanei (Saccenti et al., 2014).

Inoltre, l'approccio univariato può risultare meno sensibile in presenza di effetti deboli ma distribuiti su più variabili, che tendono invece ad emergere più chiaramente con tecniche multivariate capaci di cogliere pattern complessi. Per questo motivo, l'integrazione tra analisi univariata e multivariata è considerata una strategia solida e informativa nell'analisi di dataset ad alta dimensionalità, come quelli ottenuti tramite spettroscopia NMR (Saccenti et al., 2014).

Quando l'ANOVA rileva differenze significative tra gruppi, è prassi comune applicare test post-hoc per identificare con precisione le coppie di gruppi tra cui si osservano differenze. Tra i test più utilizzati in ambito metabolomico e agroalimentare vi è il Tukey's Honest Significant Difference (HSD), che confronta tutte le combinazioni tra gruppi controllando l'errore di tipo I associato ai confronti multipli (Sokal & Rohlf, 2013). Questo approccio risulta particolarmente utile negli studi di differenziazione geografica o varietale, o nella valutazione di effetti indotti da trattamenti tecnologici. In combinazione con l'ANOVA, il test di Tukey consente di individuare in modo mirato le variabili – come specifici bucket NMR – responsabili delle differenze tra gruppi, supportando così l'identificazione di potenziali biomarcatori discriminanti.

2.3.4 Pre-trattamento dei dati NMR

Un aspetto cruciale per garantire l'affidabilità dei risultati nell'analisi chemometrica dei dati NMR è rappresentato dalla fase di pre-trattamento, che precede le elaborazioni statistiche multivariate. Tecniche come la PCA e la PLS risultano particolarmente sensibili alla qualità, alla scala e alla coerenza dei dati in input; pertanto, è essenziale adottare procedure standardizzate in grado di ridurre le fonti di variabilità non biologicamente rilevanti.

Le operazioni di pre-processing includono generalmente la correzione della baseline, la correzione di fase e l'allineamento dei picchi. Quest'ultimo è particolarmente importante per compensare le lievi variazioni di *chemical shift* che possono verificarsi tra campioni diversi, e che, se non corrette, rischierebbero di frammentare le informazioni relative a uno stesso metabolita su bucket differenti. Altrettanto rilevante è la normalizzazione dell'intensità spettrale, che può essere eseguita rispetto all'area totale dello spettro (*total area normalization*) oppure utilizzando un riferimento interno noto, con l'obiettivo di ridurre l'impatto di differenze sistematiche di concentrazione o volume tra i campioni (Craig et al., 2006).

Una fase particolarmente delicata riguarda la scalatura delle variabili, ovvero il modo in cui ciascun bucket viene pesato prima dell'analisi multivariata. La scalatura Unit Variance (UV), che consiste nel dividere ogni variabile per la propria deviazione standard, attribuisce lo stesso peso a tutte le variabili, valorizzando anche i segnali a bassa intensità. Tuttavia, tale approccio può aumentare la sensibilità del modello al rumore. La scalatura Pareto, che prevede la divisione per la radice quadrata della deviazione standard, rappresenta un compromesso spesso preferito in metabolomica, in quanto bilancia l'influenza dei segnali forti e deboli, migliorando al contempo l'interpretabilità dei risultati (Craig et al., 2006).

Le scelte effettuate in questa fase possono influenzare significativamente le componenti latenti individuate dai modelli chemometrici e, di conseguenza, l'esito dell'intera analisi. È pertanto considerata buona pratica testare diverse modalità di scalatura e valutarne l'impatto sui pattern metabolici osservati, al fine di garantire la robustezza e la riproducibilità delle conclusioni (van den Berg et al., 2006).

2.3.5 Applicazioni in Metabolomica Alimentare

Numerosi studi presenti in letteratura evidenziano come l'integrazione tra metabolomica basata su spettroscopia NMR e analisi chemometrica mediante tecniche statistiche quali PCA, HCA, PLS-DA, OPLS-DA e ANOVA rappresenti un approccio efficace per supportare l'autenticazione, la tracciabilità e la valorizzazione dell'origine nel settore agroalimentare.

Tra i contributi rilevanti si citano i lavori di Mazzei e Piccolo (2010; 2012), condotti rispettivamente su vini Aglianico e Mozzarella di Bufala Campana, nei quali l'approccio combinato NMR–chemometria è stato applicato per valutare l'influenza del terroir e dell'origine geografica sui profili metabolici. In entrambi gli studi sono state utilizzate la PCA e la HCA per esplorare la struttura dei dati, la DA per la classificazione supervisionata e l'ANOVA per identificare i segnali spettrali significativamente discriminanti, consentendo l'individuazione di potenziali biomarcatori geografici e tecnologici.

Nella review di Consonni e Cagliani (2019), l'applicazione di PCA, PLS-DA e OPLS-DA a diverse matrici alimentari – tra cui concentrato di pomodoro, miele e zafferano – ha evidenziato l'efficacia di questi modelli nel distinguere campioni in base all'origine geografica, all'autenticità o alla presenza di adulterazioni, confermandone la versatilità in contesti applicativi differenti.

Ancora lo studio di Girelli et al. (2020), focalizzato sulla caratterizzazione varietale dell'olio extravergine di oliva, ha impiegato l'analisi multivariata basata su PCA, PLS-DA e OPLS-DA per distinguere la cultivar Coratina da altre varietà, e individuare marker metabolici discriminanti.

Più recentemente, Maestrello et al. (2022) hanno riassunto l'impiego di PCA, PLS-DA, OPLS-DA e ANOVA nel settore oleicolo, con particolare riferimento all'autenticazione e tracciabilità dell'olio extravergine di oliva. La rassegna conferma l'efficacia di questi strumenti per la caratterizzazione varietale e geografica, nonché per l'identificazione di metaboliti discriminanti legati alla cultivar o all'area di produzione.

Lo studio di Musio e Gallo (2022), focalizzato sullo zafferano, ha integrato l'uso della spettroscopia 1H-NMR con test univariati e multivariati, quali PCA, HCA, PLS-DA e OPLS-DA, riuscendo a distinguere con successo tra campioni autentici e adulterati, oltre che tra metodi di coltivazione biologica e convenzionale. L'analisi dei VIP scores e dei coefficienti di regressione ha permesso l'individuazione dei segnali spettrali maggiormente responsabili delle differenze osservate.

Ralli e Spyros (2023) hanno applicato un approccio basato su PCA, ANOVA e OPLS-DA per discriminare campioni di formaggio graviera sulla base della provenienza geografica (Creta vs. altre regioni), identificando metaboliti caratteristici associati all'origine.

Infine, Bambina et al. (2024) hanno proposto un workflow integrato basato su PCA, HCA, PLS-DA e ANOVA per analizzare vini siciliani, distinguendo varietà e aree di origine con modelli altamente predittivi e robusti. L'analisi ha portato all'identificazione di marker metabolici affidabili, validati sia con approcci multivariati che univariati.

In tutti gli studi finora citati, l'integrazione tra spettroscopia NMR e metodi statistici classici ha dimostrato di essere efficace per la classificazione dei campioni, l'individuazione di segnali discriminanti e la proposta di biomarcatori rilevanti per la tracciabilità e l'autenticazione dei prodotti agroalimentari.

Tuttavia, gli approcci statistici tradizionali presentano alcune limitazioni intrinseche. Essi si fondano spesso su assunzioni stringenti – come la normalità della distribuzione o l'indipendenza tra variabili – la cui violazione può compromettere la validità dei risultati. Inoltre, la presenza di outlier o di rumore nei dati può influenzare significativamente l'esito delle analisi, generando interpretazioni potenzialmente fuorvianti. L'interpretazione dei risultati richiede competenze avanzate, rendendo questi strumenti meno accessibili a operatori non specializzati. A ciò si aggiunge il potenziale rischio di bias soggettivi o di errori nella selezione delle variabili rilevanti. Un'ulteriore criticità riguarda infine la gestione dei dataset ad alta dimensionalità, come quelli tipici della metabolomica, per i quali i metodi classici possono risultare meno efficaci nel modellare relazioni complesse o non lineari.

Alla luce di queste limitazioni, emerge la necessità di adottare approcci più automatizzati, flessibili e robusti. In questo contesto, le tecniche basate su machine learning rappresentano una promettente evoluzione, poiché permettono di ridurre l'intervento umano, aumentare l'oggettività dell'analisi e migliorare le performance predittive in presenza di dataset complessi e ad alta dimensionalità.

2.4 Metodi Chemometrici AI-Driven in metabolomica alimentare

Negli ultimi anni, il settore della metabolomica ha conosciuto un'evoluzione significativa, registrando un crescente interesse verso l'integrazione tra le tecniche chemiometriche tradizionali — fondate principalmente su approcci statistici multivariati — e le metodologie avanzate basate sull'intelligenza artificiale (AI), quali gli algoritmi di apprendimento automatico (machine learning, ML), incluse le reti neurali artificiali (Artificial Neural Networks, ANN) (Aleixandre-Tudo et al., 2022; Corsaro et al., 2022; Chi et al., 2024). L'elevata dimensionalità e complessità dei dati metabolomici, tipicamente generati attraverso tecniche spettroscopiche come la NMR, rende infatti tali approcci particolarmente efficaci, in quanto migliorano la qualità dell'analisi, l'interpretazione biologica e la capacità predittiva complessiva dei modelli (Chi et al., 2024).

Nel contesto agroalimentare, numerosi studi recenti confermano l'efficacia dell'utilizzo di algoritmi AI in metabolomica per compiti di classificazione, regressione e selezione di variabili rilevanti per l'autenticazione e la prevenzione delle frodi (García-Pérez et al., 2024). In particolare, la combinazione di un'analisi esplorativa preliminare — svolta con metodi statistici classici, come PCA o clustering — con un'analisi predittiva basata su modelli di machine learning consente di costruire framework analitici ad alte prestazioni, applicabili al controllo di qualità e alla tutela dell'autenticità dei prodotti. Le soluzioni AI-driven permettono inoltre una significativa riduzione dei tempi analitici e delle risorse sperimentali, rendendo più efficiente l'analisi di sistemi alimentari complessi e contribuendo al rafforzamento della sicurezza e della sostenibilità della filiera (Corsaro et al., 2022; García-Pérez et al., 2024).

Le potenzialità di questi approcci nella gestione del fingerprint metabolico si rivelano particolarmente utili per la discriminazione di parametri chiave quali l'origine geografica, la varietà, il livello di freschezza e lo stadio di trasformazione del prodotto. L'elevata sensibilità mostrata dagli algoritmi AI consente infatti di rilevare variazioni metaboliche sottili, spesso non individuabili mediante i metodi statistici tradizionali, rendendoli strumenti strategici per il miglioramento della tracciabilità e del controllo dell'integrità alimentare (Kuhn et al., 2024).

Sebbene siano ancora presenti margini di miglioramento — in particolare per quanto riguarda la standardizzazione dei modelli, la gestione della riduzione dimensionale e la generalizzabilità dei risultati (Corsaro et al., 2022) — l'integrazione tra AI e metabolomica rappresenta oggi un'opportunità di grande valore per il settore agroalimentare. Tali metodologie non solo rafforzano i sistemi di certificazione della qualità e dell'origine, ma supportano anche approcci

innovativi orientati alla sostenibilità, alla sicurezza alimentare e alla valorizzazione delle produzioni locali (Antonicelli et al., 2021, García-Pérez et al., 2024).

2.4.1 Il machine learning: principi generali e principali modelli

Il *machine learning* (ML) rappresenta una branca fondamentale dell'intelligenza artificiale e si configura come un insieme di metodi computazionali capaci di apprendere automaticamente dai dati, migliorando progressivamente le proprie prestazioni nell'esecuzione di un compito specifico, senza la necessità di una programmazione esplicita.

L'applicazione del ML si è rapidamente estesa a numerosi ambiti scientifici, trovando recente impiego anche nel settore della metabolomica, in particolare in combinazione con la spettroscopia NMR (Sobolev et al., 2022). Questo sviluppo, pur favorito dall'evoluzione delle capacità computazionali, si configura anche come una risposta concreta all'esigenza di analizzare dati complessi e ad alta dimensionalità, come quelli spettroscopici, superando i limiti intrinseci dei metodi statistici convenzionali (Chi et al., 2024).

Il principio cardine del ML risiede nella costruzione di modelli matematici in grado di individuare pattern e regolarità all'interno dei dati, con l'obiettivo di generalizzare tali conoscenze per applicarle a nuovi casi. I dati impiegati nei modelli ML sono costituiti da osservazioni descritte da variabili (dette *feature*), che possono essere di tipo nominale, binario, ordinale o numerico. Il processo di apprendimento si articola generalmente in due fasi principali: l'addestramento, durante il quale il modello elabora un insieme di dati di input (con o senza etichette), e la validazione (o test), in cui le prestazioni del modello vengono valutate mediante metriche appropriate. A tal fine, si impiegano modelli matematici e statistici avanzati per stimare, tra gli altri, l'accuratezza, la sensibilità, la specificità e l'errore di predizione. Al termine della fase di apprendimento, il modello è in grado di classificare, predire o raggruppare nuove osservazioni sulla base dell'esperienza acquisita (Liakos et al., 2018).

Le principali modalità di apprendimento si distinguono in supervisionato, non supervisionato e per rinforzo. Nell'apprendimento supervisionato, il modello apprende da dati etichettati (coppie input-output), ed è particolarmente indicato per compiti di classificazione o di regressione. Nell'apprendimento non supervisionato, invece, il modello analizza dati privi di etichette, con l'obiettivo di esplorare la struttura latente e individuare pattern, anomalie o cluster. L'apprendimento per rinforzo rappresenta infine un paradigma alternativo, in cui il modello interagisce con un ambiente dinamico e ottimizza le proprie decisioni sulla base di un sistema di ricompense e penalità.

Uno strumento trasversale ai diversi approcci è la riduzione della dimensionalità (*dimensionality reduction*), essenziale per semplificare dataset complessi, come quelli generati dalla spettroscopia NMR, preservando al contempo l'informazione più rilevante. In tale ambito possono essere impiegate preliminarmente alcune tecniche chemiometriche per l'individuazione delle componenti latenti, come PCA o PLS-DA, al fine di migliorare l'efficienza e la robustezza dei modelli automatizzati (Liakos et al., 2018).

L'implementazione concreta del ML si realizza attraverso una vasta gamma di algoritmi, ciascuno caratterizzato da specifiche strategie di apprendimento, struttura e ambiti applicativi. Nelle sezioni seguenti vengono presentati alcuni modelli selezionati in virtù del loro utilizzo in diversi studi condotti in ambito agroalimentare (Corsaro et al., 2022, Botero-Valencia et al., 2025). Essi trovano applicazione sia per finalità di classificazione (es discriminazione in base alla varietà geografica o botanica o al metodo di trasformazione) che di regressione (es. stima del contenuto metabolico).

2.4.2 Modelli di regressione

I modelli di regressione appartengono all'ambito dell'apprendimento supervisionato, e permettono di stimare una variabile di output continua a partire da variabili indipendenti note. Gli algoritmi più semplici includono la regressione lineare, la regressione logistica (utilizzata per compiti di classificazione) e la regressione stepwise (Liakos et al., 2018).

2.4.2.1 Regressione Logistica (Logistic Regression)

La regressione logistica è uno dei modelli statistici più utilizzati per compiti di classificazione binaria. A differenza della regressione lineare, che fornisce un output continuo, questo algoritmo stima la probabilità che un'osservazione appartenga a una determinata classe, sulla base delle variabili indipendenti. Tale probabilità è calcolata attraverso una funzione logistica (sigmoide), che trasforma una combinazione lineare delle feature in un valore compreso tra 0 e 1. Se il valore supera una soglia predefinita (di norma pari a 0.5), l'osservazione viene assegnata alla classe positiva; in caso contrario, alla classe negativa (Hosmer et al., 2013).

Da un punto di vista teorico, la regressione logistica è un modello discriminativo: stima direttamente la probabilità condizionata dell'output dato un insieme di input. I coefficienti associati alle variabili esplicative vengono stimati tramite massima verosimiglianza, con l'obiettivo di massimizzare la probabilità di osservare i dati del training set. Il modello può essere esteso a scenari multiclasse, ad esempio attraverso strategie one-vs-rest (James et al., 2021).

Tra i principali vantaggi vi sono la semplicità, l'efficienza computazionale e l'elevata interpretabilità. I coefficienti stimati permettono di comprendere l'effetto diretto (positivo o negativo) di ciascuna variabile sulla probabilità di appartenenza alla classe positiva, offrendo uno strumento utile per la selezione di biomarcatori.
Tuttavia, il modello presenta anche alcune limitazioni. La principale riguarda l'assunzione di una relazione lineare tra le feature e l'output probabilistico del modello (*log-odds*), che può risultare inadeguata in presenza di strutture decisionali complesse o non lineari. In tali casi, è necessario introdurre manualmente termini polinomiali o di interazione, oppure adottare modelli più flessibili. Inoltre, la regressione logistica è sensibile alla multicollinearità tra le variabili e ai valori anomali, che possono compromettere la stima dei parametri e ridurre l'affidabilità del modello, soprattutto in presenza di dataset piccoli o sbilanciati (Lukman et al., 2025).

2.4.3 Modelli bayesiani

I modelli bayesiani costituiscono una famiglia di modelli probabilistici basati sull'inferenza bayesiana. Anch'essi rientrano nell'apprendimento supervisionato e vengono utilizzati sia per classificazione sia per regressione. Tra gli algoritmi più diffusi si annovera il Naive Bayes, nelle sue varianti gaussiana e multinomiale, utilizzato per stimare la probabilità di appartenenza a una classe sulla base della distribuzione condizionale delle feature (Liakos et al., 2018).

2.4.3.1 Naive Bayes

Il classificatore Naive Bayes è un modello di apprendimento supervisionato appartenente alla famiglia dei metodi generativi. Il suo funzionamento si basa sul teorema di Bayes, secondo cui è possibile calcolare la probabilità che un'osservazione appartenga a una determinata classe in funzione delle caratteristiche osservate. La peculiarità del modello risiede nell'assunzione di indipendenza condizionata: si assume, cioè, che ogni variabile (feature) sia indipendente dalle altre, a parità di classe. Questa ipotesi, sebbene raramente soddisfatta nei dati reali, consente una notevole semplificazione computazionale e un'implementazione molto rapida.
A seconda della natura delle variabili, esistono diverse varianti del classificatore Naive Bayes: il Naive Bayes Gaussiano, adatto a feature continue, assume una distribuzione normale per ciascuna variabile condizionata alla classe; le versioni multinomiali e bernoulliane sono invece più indicate per dati discreti o binari, come conteggi o presenze/assenze.
Tra i principali punti di forza del modello si annoverano l'elevata rapidità di addestramento, la robustezza in condizioni di dati scarsamente rappresentati e la sua naturale estensione ai

problemi di classificazione multi-classe, senza la necessità di ricorrere a strategie complesse di ricodifica (Blanquero et al., 2021).

Tuttavia, il principale limite del classificatore risiede nella semplificazione dell'indipendenza tra le variabili, ipotesi difficilmente rispettata in ambiti ad alta dimensionalità come l'NMR, dove i segnali spettrali tendono a essere fortemente correlati. In tali casi, il modello può sovra- o sottostimare le probabilità predette, con un impatto negativo sulla classificazione. Inoltre, le probabilità calcolate tendono ad essere non ben calibrate, cioè molto vicine a 0 o 1, il che può compromettere decisioni basate su soglie di rischio. Un ulteriore limite è rappresentato dal problema della zero-frequency: se una combinazione tra valore e classe non è mai comparsa nel training set, la probabilità stimata sarà pari a zero. Questo può essere corretto con tecniche di smoothing, come la correzione di Laplace (Blanquero et al., 2021).

2.4.4 Modelli basati sugli esempi (Instance-Based Models, IBM)

Questi modelli si fondano sulla memorizzazione delle istanze del set di addestramento e realizzano previsioni confrontando direttamente nuovi esempi con quelli già osservati, senza generalizzare attraverso una funzione astratta. Il processo decisionale si basa quindi su confronti puntuali, tipicamente condotti tramite misure di distanza. Un limite di questi modelli è rappresentato dalla loro elevata complessità computazionale al crescere della dimensione del dataset. Uno degli algoritmi più rappresentativi è il k-nearest neighbor (k-NN) (Liakos et al., 2018).

2.4.4.1 k-Nearest Neighbors (k-NN)

k-Nearest Neighbors (k-NN) è un algoritmo di apprendimento supervisionato basato sulla similarità tra osservazioni. Introdotto da Cover e Hart (1967), si tratta di un metodo di tipo *instance-based*, in cui non viene costruito un modello vero e proprio: per classificare o stimare un nuovo campione, il sistema si limita a confrontarlo con i k esempi più simili nel dataset di addestramento, misurando la distanza nello spazio delle variabili (solitamente con la distanza euclidea). Nei dati NMR, ad esempio, tali variabili corrispondono ai bucket spettrali opportunamente normalizzati e scalati.

L'assegnazione ad una determinata classe avviene in funzione della minima distanza calcolata. Il funzionamento è semplice: se un campione sconosciuto ha un profilo NMR simile a quello di più campioni di olio extravergine autentico, il modello lo assegnerà a quella classe.

Questo approccio intuitivo non richiede addestramento esplicito e non impone ipotesi sulla forma della funzione decisionale, risultando particolarmente utile come *baseline* per confrontare altri modelli (Altman, 1992).

Tuttavia, il k-NN presenta limiti noti, soprattutto con dati ad alta dimensionalità come quelli NMR. In spazi con molte variabili, le distanze tra i punti tendono a diventare simili, rendendo difficile distinguere i "veri" vicini (Beyer et al., 1999). Inoltre, è computazionalmente costoso su dataset estesi, poiché ogni predizione richiede il confronto con tutti i campioni del training set.

Per mitigare questi problemi, possono essere impiegate tecniche di riduzione della dimensionalità come la PCA (Liakos et al., 2018). Infine, la scelta del parametro k è cruciale: valori troppo bassi rendono il modello instabile, mentre valori troppo alti possono diluire le informazioni utili, compromettendo la precisione della classificazione (Cover & Hart, 1967).

2.4.5 Alberi decisionali (Decision Trees, DT)

Gli alberi decisionali sono tra i modelli di apprendimento supervisionato più diffusi, utilizzati sia per compiti di classificazione che di regressione. Nel caso della classificazione, il modello costruisce una struttura ad albero, in cui ogni nodo interno rappresenta una condizione basata sul valore di una variabile (feature), e ogni ramo corrisponde a un possibile esito. Le foglie finali indicano la classe predetta per l'osservazione (Liakos et al., 2018).

L'addestramento avviene attraverso una partizione ricorsiva del dataset, suddividendo progressivamente i dati in gruppi omogenei rispetto alla variabile target. La scelta delle suddivisioni si basa su misure come l'indice di Gini o l'entropia, con l'obiettivo di migliorare la purezza delle classi nei nodi risultanti (Quinlan, 1986; Breiman et al., 1984). I principali algoritmi per la costruzione di alberi includono ID3, C4.5 e CART (Classification and Regression Trees).

Uno dei principali vantaggi degli alberi decisionali è la loro trasparenza: la sequenza di regole if–then è facile da interpretare e comunicare, anche in contesti non specialistici. Inoltre, questi modelli sono in grado di gestire sia variabili continue che categoriche, richiedono una preparazione preliminare dei dati relativamente contenuta, possono modellare relazioni non lineari e risultano robusti anche in presenza di valori mancanti o dati rumorosi (Kuhn & Johnson, 2013).

Tuttavia, un albero decisionale singolo può risultare instabile: piccole variazioni nei dati possono produrre strutture molto diverse. Inoltre, se lasciato crescere senza limiti, tende a sovradattarsi (overfitting) ai dati di addestramento, perdendo capacità predittiva su dati nuovi.

Per ridurre questo rischio si adottano strategie come la potatura (pruning), il controllo della profondità massima dell'albero, oppure si impongono soglie minime per il numero di campioni nei nodi terminali (Hastie et al., 2009).
Proprio per superare questi limiti, sono stati sviluppati modelli ensemble, come Random Forest e XGBoost, che combinano più alberi in modo da migliorare la stabilità e le prestazioni complessive del sistema predittivo (Breiman, 2001).

2.4.6 Modelli di ensemble learning (EL)

I modelli di ensemble learning (EL) mirano a migliorare le performance predittive combinando più modelli di base (*base learners*) in una struttura aggregata. Ogni modello dell'ensemble rappresenta una singola ipotesi, e la loro combinazione consente di ottenere risultati più accurati e stabili, soprattutto quando i singoli modelli sono tra loro diversificati. Gli alberi decisionali sono comunemente utilizzati come base learner in molti algoritmi ensemble, come nel caso del Random Forest.
Due approcci principali all'interno dell'ensemble learning sono il bagging (*bootstrap aggregating*) e il boosting. Il bagging costruisce diversi modelli in parallelo, ciascuno addestrato su un sottoinsieme casuale del dataset (campionamento con ripetizione), e combina le predizioni per ridurre la varianza e migliorare la stabilità del sistema. Il boosting, invece, costruisce i modelli in modo sequenziale, assegnando maggiore peso agli errori commessi dai modelli precedenti, così da correggerli progressivamente e ridurre il bias. Tra i metodi più noti rientrano il Random Forest per il bagging, e algoritmi come AdaBoost, Gradient Boosting e XGBoost per il boosting (Liakos et al., 2018).

2.4.6.1 Random Forest

Il Random Forest (RF) è un algoritmo di *ensemble learning* che combina numerosi alberi decisionali, ciascuno addestrato su un sottocampione diverso del dataset originale. Ogni albero produce una predizione indipendente — sotto forma di classificazione o regressione — basata su una serie di regole decisionali *if–then*, costruite considerando un sottoinsieme casuale delle variabili disponibili (ad esempio, i bucket spettrali nei dati NMR).
La predizione finale del modello si ottiene aggregando le risposte di tutti gli alberi: nel caso della classificazione si adotta il criterio della maggioranza (*majority voting*), mentre nella regressione si calcola la media delle predizioni. Questo meccanismo consente di ridurre l'impatto del rumore o di alberi particolarmente deboli, aumentando così la robustezza e la capacità di generalizzazione del modello complessivo (Breiman, 2001).

Un aspetto distintivo del Random Forest è la presenza di casualità controllata sia nella selezione dei dati sia delle variabili. Ogni albero viene addestrato su un campione casuale (*bootstrapped*) del dataset originale, e a ogni nodo decisionale viene valutato solo un sottoinsieme casuale di variabili per determinare la migliore suddivisione. Questo approccio riduce la correlazione tra gli alberi e migliora l'effetto di aggregazione dell'ensemble, riducendo il rischio di overfitting, particolarmente rilevante nei dataset ad alta dimensionalità come quelli derivanti da spettroscopia NMR (Cutler et al., 2007).

2.4.6.2 XGBoost (Extreme Gradient Boosting)

XGBoost è un algoritmo di apprendimento supervisionato appartenente alla famiglia del *boosting* basato su alberi decisionali. A differenza del metodo *bagging* utilizzato da Random Forest, in cui i modelli sono costruiti in parallelo su sottocampioni diversi, nel boosting gli alberi vengono costruiti in sequenza: ciascun nuovo albero è addestrato per correggere gli errori commessi dai modelli precedenti.

In particolare, XGBoost implementa una versione ottimizzata del gradient boosting, costruendo ad ogni iterazione un nuovo albero che minimizza una funzione di perdita differenziabile rispetto ai residui delle predizioni correnti. Il risultato è un ensemble di alberi deboli, ciascuno poco profondo, ma in grado di generare un predittore potente e accurato quando combinati tra loro (Chen & Guestrin, 2016).

Tra i principali punti di forza di XGBoost vi sono la velocità di addestramento, l'elevata accuratezza predittiva, e la capacità di gestire dataset ad alta dimensionalità, anche in presenza di molte variabili irrilevanti, grazie alla selezione intrinseca delle feature. Il modello calcola automaticamente l'importanza delle variabili (feature importance), rendendolo utile anche per individuare potenziali biomarcatori nei dati NMR (Lundberg et al., 2020).

XGBoost include molte funzionalità avanzate come la regolarizzazione L1/L2, la gestione dei dati mancanti, la potatura preventiva degli alberi, e il supporto alla parallelizzazione, che ne hanno favorito l'adozione diffusa anche in competizioni internazionali di data science (Chen & Guestrin, 2016). Tuttavia, la flessibilità del modello comporta anche una maggiore complessità nella fase di tuning: è necessario selezionare accuratamente parametri come il numero di alberi, la profondità massima, il *learning rate*, la frazione di dati usati a ogni iterazione (*subsample*) e i coefficienti di regolarizzazione. Se non adeguatamente configurato, il modello può andare incontro a overfitting, soprattutto in presenza di dataset di piccole dimensioni o classi sbilanciate (Probst et al., 2019; Kuhn & Johnson, 2013).

2.4.7 Support Vector Machines (SVM)

Le Support Vector Machines (SVM) sono algoritmi di apprendimento supervisionato introdotti per la classificazione binaria, ma oggi ampiamente estesi anche a problemi di regressione (Support Vector Regression, SVR) e, in alcuni casi, di clustering. Il principio di base delle SVM consiste nell'individuare un iperpiano ottimale che separi nel modo più netto possibile due classi, massimizzando la distanza (o margine) tra il piano stesso e i campioni più vicini di ciascuna classe, detti vettori di supporto (Cortes & Vapnik, 1995, Liakos et al., 2018).

Quando i dati non sono linearmente separabili nello spazio delle variabili originali, le SVM possono sfruttare il kernel trick, una tecnica matematica che consente di proiettare i dati in uno spazio a dimensionalità superiore, dove è possibile realizzare una separazione lineare (Hsu et al., 2010; Kuhn & Johnson, 2013). Tra i kernel più comunemente impiegati si annoverano quelli lineare, polinomiale e Radial Basis Function (RBF), ciascuno dei quali permette di modellare differenti tipi di relazioni complesse tra le variabili (Corsaro et al., 2022).

Nonostante l'efficacia del metodo, l'SVM presenta anche alcuni limiti. In particolare, risulta generalmente meno interpretabile rispetto ad altri modelli (es. Random Forest), in quanto non fornisce direttamente una stima intuitiva dell'importanza delle variabili (Cortez & Embrechts, 2013), sebbene esistano strategie per estrarre pesi nei modelli lineari. Inoltre, le sue performance dipendono fortemente dalla scelta dei parametri di tuning, quali il tipo di kernel, il parametro di penalizzazione *C* e, nel caso di kernel non lineari, il parametro *gamma* (Hsu et al., 2010; Kuhn & Johnson, 2013). Per questo motivo è necessaria una fase di ottimizzazione mediante validazione incrociata. Un'ulteriore criticità può emergere in presenza di etichette rumorose o classi parzialmente sovrapposte, che possono compromettere l'accuratezza del modello; tuttavia, la formulazione *soft-margin* consente di tollerare piccoli errori di classificazione, migliorando così la robustezza complessiva del sistema (Cortes & Vapnik, 1995).

2.4.8 Reti Neurali Artificiali (ANN)

Le Artificial Neural Networks (ANN) sono modelli computazionali ispirati in modo semplificato al funzionamento dei neuroni biologici. Sono composte da unità elementari (nodi o neuroni artificiali) organizzate in strati interconnessi: uno di input, uno o più strati nascosti e uno strato di output. Ogni nodo riceve segnali dallo strato precedente, li combina linearmente tramite pesi numerici, applica una funzione di attivazione (ad esempio *sigmoid*, *ReLU* o *tanh*), e trasmette il risultato allo strato successivo (Liakos et al., 2018).

Le architetture a più strati con funzioni di attivazione non lineari, note come Multilayer Perceptron (MLP), sono teoricamente in grado di approssimare qualunque funzione, purché dispongano di sufficiente capacità computazionale e di un numero adeguato di dati per l'addestramento (Hornik, 1991; Corsaro et al., 2022).

Un ambito emergente è quello del deep learning, che impiega reti neurali profonde (Deep Neural Networks, DNN) con molteplici strati nascosti. Tra le architetture più note si annoverano le Convolutional Neural Networks (CNN), particolarmente efficaci nell'analisi di segnali e immagini, le Reti Neurali Ricorrenti (RNN) e le Long Short-Term Memory (LSTM), adatte alla modellazione di dati sequenziali.

Tuttavia, l'applicazione pratica del deep learning richiede dataset ampi, etichettati e ben standardizzati, condizione spesso difficile da soddisfare nei contesti agroalimentari. Per questo motivo, si tende a preferire architetture relativamente semplici, integrate con tecniche di regolarizzazione come il *dropout* o la penalizzazione dei pesi (*weight decay*), per evitare il fenomeno dell'overfitting, cioè un'eccessiva aderenza ai dati di training a scapito della capacità predittiva su dati nuovi (Srivastava et al., 2014).

Il punto di forza delle ANN risiede nella loro flessibilità modellistica: non richiedono ipotesi a priori sulla forma della relazione tra le variabili, e possono modellare interazioni complesse e non lineari nei profili metabolici. Tuttavia, un limite importante è la scarsa interpretabilità: i pesi appresi non sono facilmente riconducibili a variabili specifiche, rendendo difficile attribuire un significato biochimico diretto alle predizioni (Montavon et al., 2018; Corsaro et al., 2022).

2.4.9 Vantaggi e Limiti: Confronto tra Metodi Classici e Approcci di ML

Nell'ambito della presente ricerca, focalizzata sull'analisi di dati NMR derivanti da matrici alimentari, è possibile confrontare i metodi statistici classici con gli approcci basati sul Machine Learning (ML). Entrambi gli approcci presentano vantaggi distinti e spesso complementari, differenziandosi nella capacità di modellare la complessità dei dati, nella robustezza predittiva, nella scalabilità e nell'interpretabilità dei modelli.

Capacità di modellare relazioni complesse: Gli algoritmi di ML si distinguono per la loro abilità nel catturare relazioni non lineari e interazioni complesse tra variabili spettrali. Questi modelli sono particolarmente efficaci in presenza di segnali deboli, variabili ridondanti o correlazioni intricate, caratteristiche tipiche dei dati della Spettroscopia NMR (Kuhn & Johnson, 2013; James et al., 2021). Studi recenti hanno evidenziato che tali algoritmi superano

le performance dei modelli convenzionali in diversi contesti applicativi, inclusa la discriminazione varietale, la classificazione geografica e il rilevamento di adulterazioni (Nyitrainé Sárdy et al., 2022; Pollak et al., 2024).

Richiesta di dati e rischio di overfitting: Uno dei principali limiti degli algoritmi di apprendimento automatico (ML) è la loro elevata richiesta di dati per poter essere addestrati in modo efficace e generalizzabile. In ambito agroalimentare, la raccolta di dataset ampi e rappresentativi può risultare onerosa, specialmente quando si lavora con prodotti a denominazione d'origine, annate specifiche o varietà botaniche rare, per cui il numero di campioni disponibili può essere limitato.

In presenza di dataset di piccole dimensioni, l'elevata flessibilità di alcuni modelli avanzati – come le reti neurali o gli ensemble methods (es. Random Forest) – li rende più suscettibili al fenomeno di overfitting, ovvero la tendenza del modello ad adattarsi in modo eccessivo ai dati di addestramento, includendo anche il rumore specifico e perdendo capacità di generalizzazione (Kuhn et al., 2013).

I metodi statistici classici come PCA, PLS-DA risultano in genere più parsimoniosi, poiché proiettano i dati su un numero limitato di componenti latenti. Questo tipo di trasformazione agisce come una regolarizzazione implicita, contribuendo a mitigare l'overfitting e rendendo i modelli più stabili, soprattutto in presenza di dati ad alta dimensionalità e basso numero di campioni.

Per i modelli ML, evitare l'overfitting richiede l'impiego di strategie di validazione rigorose, come la k-fold cross-validation ripetuta, la validazione su set esterni e, ove possibile, test di permutazione o double cross-validation (Hastie et al., 2009, Kuhn et al., 2013).

Interpretabilità dei risultati: Nell'ambito agroalimentare, e in particolare nello studio dell'autenticità e della tracciabilità dei prodotti, l'interpretabilità dei modelli predittivi riveste un ruolo strategico. Non è sufficiente determinare se un campione appartenga o meno a una determinata classe (es. prodotto autentico vs contraffatto); è infatti cruciale comprendere *perché* è stata assegnata tale etichetta. L'identificazione di marker chimici discriminanti – come metaboliti caratteristici di una specifica origine geografica o varietale – può avere importanti implicazioni sia normative che tecnologiche e commerciali. In tal senso, i metodi chemometrici classici, come PCA e PLS-DA, sono da tempo impiegati per la loro trasparenza interpretativa. La PCA consente, ad esempio, di visualizzare i loading plots, che identificano le variabili (es. segnali NMR o bucket) maggiormente responsabili delle separazioni osservate lungo le componenti principali. Nei modelli supervisionati come la PLS-DA è possibile ottenere VIP

scores (Variable Importance in Projection) e coefficienti di regressione che evidenziano le feature più determinanti nella classificazione (Sobolev et al., 2022).

Al contrario, i modelli avanzati di machine learning, come le Support Vector Machines (SVM) o le Artificial Neural Networks (ANN), sono spesso definiti *black-box models* per la bassa leggibilità interna del processo decisionale. Un modello SVM con kernel non lineare costruisce ipersuperfici complesse, mentre una rete neurale distribuisce l'informazione attraverso migliaia di pesi sinaptici, rendendo difficile attribuire importanza alle singole variabili. Tuttavia, non tutti i modelli AI sono completamente opachi: algoritmi come Random Forest e XGBoost, pur essendo non lineari, permettono di stimare l'importanza delle variabili tramite metodi come il "Mean Decrease in Accuracy" (MDA) o il Gain associato a ogni nodo decisionale (Hou et al., 2024; Maestrello et al., 2024).

In generale, i modelli statistici classici offrono una lettura lineare, intuitiva e comunicabile dei risultati, rendendoli ideali quando l'obiettivo è l'individuazione di biomarcatori affidabili o la comprensione dei meccanismi biochimici alla base della classificazione. I modelli ML, al contrario, risultano particolarmente vantaggiosi in scenari ad alta complessità e su larga scala, in cui è richiesta massima accuratezza predittiva. In tali casi, può essere utile affiancare al modello predittivo un'analisi interpretativa parallela, basata su PCA, VIP score o analisi univariata (ANOVA, Tukey HSD), al fine di garantire una comprensione completa del sistema e favorire la trasferibilità dei risultati (Bifarin, 2023).

2.4.10 Metriche di Performance e Validazione dei Modelli

Per valutare quantitativamente le prestazioni dei modelli di machine learning (ML), si ricorre a diverse metriche, la cui scelta dipende dalla natura del compito predittivo — classificazione o regressione — e dalla distribuzione delle classi nel dataset. Nel contesto della classificazione, le metriche più comunemente adottate sono l'accuratezza, la precisione, il richiamo (recall) e l'F1-score.

Accuratezza (Accuracy): rappresenta la frazione di osservazioni correttamente classificate sul totale e costituisce una metrica intuitiva e di facile interpretazione. In ambito agroalimentare, ad esempio, può indicare la percentuale di campioni correttamente assegnati alla regione di origine sulla base del profilo spettrale NMR. Tuttavia, in presenza di classi sbilanciate, l'accuratezza può risultare fuorviante: un modello che predice sempre la classe prevalente può ottenere un'accuratezza elevata anche senza reale capacità discriminante (Saito & Rehmsmeier, 2015).

Per ovviare a questa limitazione, è prassi accompagnare l'accuratezza con metriche che tengano conto della distribuzione delle classi: come la precisione, la sensibilità (recall) e la F1-score.
Nel caso di classificazione multiclasse, si può calcolare l'accuratezza bilanciata (balanced accuracy), ovvero la media dell'accuratezza ottenuta su ciascuna classe, particolarmente indicata quando le classi non sono equi-distribuite. Questa metrica consente di ridurre il bias verso le classi maggioritarie, migliorando la rappresentatività delle prestazioni globali del modello (Brodersen et al., 2010).

Precisione, Recall e F1-score: oltre all'accuratezza, metriche come precisione, recall (o sensibilità) e F1-score rivestono un ruolo fondamentale nella valutazione delle prestazioni di un modello classificatore, soprattutto in presenza di dataset sbilanciati o quando gli errori di classificazione hanno un impatto differenziale in base alla classe coinvolta. La Precisione (Precision): indica la proporzione di predizioni corrette tra tutte quelle etichettate come positive. Una precisione elevata implica che il modello commette pochi falsi positivi. Il Richiamo (Recall), o sensibilità, misura la frazione di veri positivi identificati rispetto al totale reale di elementi positivi. Un alto richiamo indica che il modello è efficace nel non lasciarsi sfuggire casi rilevanti. La metrica F1-score rappresenta la media armonica tra precisione e recall. È particolarmente utile quando è necessario bilanciare l'accuratezza nella rilevazione con l'affidabilità delle predizioni (Sokolova & Lapalme, 2009; Saito & Rehmsmeier, 2015).
Nel contesto della metabolomica da spettroscopia NMR, queste metriche sono frequentemente utilizzate per valutare le prestazioni di modelli di classificazione automatica basati su dati spettrali ad alta dimensionalità (Nyitrainé-Sárdy et al., 2022).

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): nel contesto della classificazione binaria — o in problemi multiclasse gestiti con approcci "one-vs-rest" — una metrica particolarmente apprezzata per valutare le performance dei modelli è l'AUC-ROC, ovvero l'area sotto la curva ROC (Receiver Operating Characteristic). Questa curva mette in relazione la sensibilità (recall) con 1-specificità (cioè il tasso di falsi positivi), al variare della soglia decisionale utilizzata per la classificazione.
L'AUC fornisce una misura sintetica della capacità del modello di discriminare correttamente tra le classi: un valore pari a 1.0 indica una classificazione perfetta, mentre un valore di 0.5 equivale a un comportamento casuale. Valori superiori a 0.8 sono generalmente considerati buoni, anche se il contesto applicativo può influenzare l'interpretazione (Bradley, 1997).
Questa metrica ha il vantaggio di essere indipendente dalla soglia di decisione, ed è quindi particolarmente utile in scenari dove il bilanciamento tra le classi è marcato o quando

l’ottimizzazione della soglia non è immediata, come spesso avviene nelle applicazioni agroalimentari su dati NMR.

Nel campo della chemometria applicata alla spettroscopia NMR, l’AUC viene frequentemente utilizzata per confrontare modelli di classificazione supervisionata (Nyitrainé Sárdy et al., 2022).

2.4.11 Strategie di validazione

Indipendentemente dalla metrica utilizzata per la valutazione delle prestazioni di un modello, è fondamentale che tale valutazione venga effettuata su dati non utilizzati nella fase di addestramento, al fine di ottenere una stima realistica della capacità di generalizzazione. Questo passaggio è cruciale per evitare sovrastime delle performance dovute all’overfitting, ossia al rischio che il modello si adatti eccessivamente alle specificità del dataset di training, perdendo capacità predittiva su nuovi dati.

Validazione incrociata (k-fold cross-validation): Una delle tecniche più diffuse nell’ambito del machine learning e della chemometria è la validazione incrociata k-fold. Tale strategia prevede la suddivisione del dataset in k sottoinsiemi (fold), utilizzando in ciascuna iterazione k–1 fold per l'addestramento e il fold rimanente per la validazione. Il processo viene ripetuto k volte, in modo che ogni fold venga utilizzato una sola volta come set di test. I risultati ottenuti dalle diverse iterazioni vengono infine aggregati per fornire una stima complessiva e robusta delle performance del modello su dati non visti. In metabolomica alimentare, dove il numero di campioni è spesso limitato, è comune l’utilizzo di configurazioni 5-fold o 10-fold, spesso con ripetizioni multiple (repeated CV) per ridurre la varianza delle stime e garantire una maggiore stabilità dei risultati (Vu et al., 2019). La validazione incrociata si rivela particolarmente utile anche durante la fase di ottimizzazione degli iperparametri: ad esempio, per selezionare il numero ottimale di alberi in una Random Forest, oppure il valore del parametro di penalità *C* e del *kernel* in un modello SVM, in funzione dell’accuratezza o dell’F1-score mediamente ottenuti.

In presenza di dataset sbilanciati, è consigliata l’adozione di una validazione incrociata stratificata, che assicura la conservazione della proporzione tra le classi in ciascun fold, riducendo il rischio di bias nelle stime di performance (Kuhn & Johnson, 2013).

Set di test esterno: L'uso di un set di test esterno è fondamentale per valutare la capacità di un modello di generalizzare a nuovi dati non visti durante l'addestramento o la validazione interna. Questo approccio è particolarmente cruciale in applicazioni pratiche come il controllo qualità

e l'autenticazione alimentare, dove i dati futuri possono differire significativamente da quelli utilizzati per sviluppare il modello.

Ad esempio, uno studio sull'autenticazione delle carote biologiche ha utilizzato campioni provenienti da diverse annate (2005–2008) come set di test esterni per valutare la robustezza del modello, evidenziando l'importanza di includere variazioni temporali nei dati di test (Cubero-Leon e t al., 2018).

In un altro studio, è stato sviluppato un modello di machine learning per autenticare il sistema di allevamento delle galline ovaiole basato su dati NMR di oltre 4.000 campioni di tuorlo d'uovo. Il modello è stato testato su campioni acquistati in supermercati, dimostrando la sua applicabilità in scenari reali e l'importanza di un set di test esterno per valutare la performance del modello (Bischof et al., 2024).

Bootstrapping: Il bootstrapping è una tecnica di validazione statistica che consiste nel generare un gran numero di sottocampioni casuali con sostituzione dal dataset originale. Su ciascuno di questi sottocampioni, il modello viene addestrato e testato, consentendo così di stimare in modo affidabile la variabilità delle performance del modello (es. accuratezza, AUC, etc.), di calcolare intervalli di confidenza e di valutare la stabilità delle predizioni.

Uno degli utilizzi più significativi del bootstrapping è la costruzione di bande di confidenza per le curve ROC, permettendo di quantificare la variabilità associata all'area sotto la curva (AUC) e di confrontare la robustezza di modelli alternativi. In ambienti applicativi come la classificazione di matrici alimentari mediante NMR, questa tecnica può contribuire a rafforzare la validazione statistica quando i dataset sono limitati o eterogenei (Lankham & Slaughter, 2020).

Controllo dell'overfitting: Un indicatore critico della qualità di un modello predittivo è la coerenza tra le performance ottenute sui dati di addestramento e quelle osservate su dati non visti, come nel caso della cross-validation o di un test set esterno. In presenza di overfitting, il modello raggiunge un'accuratezza elevata in fase di training, ma mostra un degrado significativo nelle prestazioni su nuovi campioni. Questo fenomeno indica che il modello ha appreso anche il rumore e le peculiarità specifiche del dataset di addestramento, compromettendo la capacità di generalizzare (Hastie et al., 2009).

Per mitigare tale rischio, si possono adottare diverse strategie, sia a livello strutturale che procedurale. Una prima misura consiste nel limitare la complessità del modello, ad esempio riducendo il numero di neuroni nei livelli nascosti di una rete neurale o controllando la profondità degli alberi decisionali, per evitare l'apprendimento di pattern non generalizzabili (Hastie et al., 2009). Un'altra tecnica efficace è la regolarizzazione: l'introduzione di

penalizzazioni L1 (Lasso) o L2 (Ridge) sui coefficienti del modello aiuta a contenere la magnitudine dei pesi, favorendo soluzioni più parsimoniose e stabili (Ng, 2004).
Ulteriori approcci includono l'early stopping, che consiste nell'interrompere l'addestramento quando le performance su un validation set iniziano a peggiorare, segnalando l'inizio del sovradattamento (Prechelt, 1998), e l'uso del dropout nei modelli neurali. Quest'ultimo prevede la disattivazione casuale di neuroni durante l'apprendimento, riducendo la dipendenza del modello da specifici percorsi predittivi e migliorando così la capacità di generalizzazione (Srivastava et al., 2014).

2.4.12 Applicazioni in Metabolomica Alimentare

Numerose evidenze tratte da letteratura scientifica confermano l'efficacia dell'integrazione tra spettroscopia NMR e tecniche avanzate di machine learning (ML) per l'autenticazione, la classificazione e la valorizzazione qualitativa dei prodotti agroalimentari.
Tra i contributi recenti si segnala lo studio di Nyitrainé-Sárdy et al. (2022), nel quale 861 campioni di vino rosso sono stati analizzati mediante spettroscopia 1H-NMR e classificati attraverso diversi algoritmi di ML (Support Vector Machine, Random Forest e Reti Neurali). I modelli hanno raggiunto accuratezze predittive fino al 95% nella discriminazione varietale e geografica, identificando metaboliti chiave (tramite valutazione della MDA) fortemente associati sia alla varietà botanica che al terroir. Lo studio ha confermato l'utilità di tale approccio per l'autenticazione e il controllo qualità dei vini.
Nel lavoro di Cui et al. (2023), l'approccio NMR–ML è stato applicato a 219 campioni di tè nero al fine di discriminare l'origine geografica, la varietà botanica e il metodo di trasformazione. Tra i modelli testati (k-Nearest Neighbors, Support Vector Machine, Random Forest), Random Forest ha ottenuto le migliori performance (accuratezza del 92.7%), consentendo l'identificazione di metaboliti discriminanti (tramite valutazione della MDA) rappresentativi del fingerprint chimico del prodotto, e dimostrando l'efficacia dell'approccio per la tracciabilità e la lotta alla contraffazione.
Uno studio analogo è quello di Hou et al. (2024), incentrato sull'autenticazione del tè Longjing, uno dei più noti e pregiati prodotti a denominazione di origine protetta (PDO) della Cina. Anche in questo caso, l'integrazione tra spettroscopia 1H-NMR e algoritmi di machine learning si è dimostrata efficace per la tracciabilità geografica. È stato sviluppato un modello Random Forest addestrato per discriminare campioni provenienti da quattro diverse aree di coltivazione. Il modello ha raggiunto un'accuratezza del 92.3% sul test set, mostrando inoltre elevata robustezza nei confronti di campioni provenienti da regioni non certificate PDO.

L’identificazione dei metaboliti discriminanti, condotta tramite valutazione della MDA, ha permesso di individuare composti chiave per la tracciabilità geografica. Lo studio sottolinea l’efficacia di questo approccio anche nel supportare la tutela delle denominazioni d’origine contro fenomeni di adulterazione e contraffazione.

Ramiro et al. (2024) hanno applicato l’integrazione tra spettroscopia NMR e algoritmi di machine learning (NMR–ML) per la classificazione di carni suine e la quantificazione dei profili lipidici. I dati NMR, acquisiti su diverse razze e tipologie di taglio, sono stati elaborati con vari algoritmi supervisionati (Support Vector Machine, Random Forest, Gradient Boosting, Naive Bayes, k-Nearest Neighbors), ottenendo accuratezze superiori al 95% per la maggior parte dei modelli e, nel caso dell’SVM, un’accuratezza del 100% sul test set. Sono stati inoltre sviluppati modelli di regressione per stimare la composizione lipidica (SFA, MUFA, PUFA), che hanno mostrato performance predittive eccellenti. I risultati confermano la solidità dell’approccio anche su matrici complesse e aprono interessanti prospettive per la realizzazione di sistemi intelligenti di autenticazione, controllo qualità e tracciabilità nel settore agroalimentare.

Nel settore lattiero-caseario, Maestrello et al. (2024) hanno proposto un approccio combinato basato su spettroscopia 1H-NMR e algoritmi di machine learning, in particolare Random Forest, per l’autenticazione del formaggio Grana Padano DOP. L’analisi ha coinvolto 144 campioni, comprendenti sia formaggi autentici DOP che prodotti concorrenti, rappresentativi di diverse tipologie e provenienze geografiche. Sono stati adottati sia un approccio untargeted, basato sull’intero spettro NMR, sia un approccio targeted, focalizzato sulle frazioni acquosa e lipidica dopo specifico pretrattamento, con quantificazione di metaboliti selezionati. Il modello Random Forest ha evidenziato ottime performance predittive in entrambi i casi (AUC > 0.93), dimostrandosi efficace anche nella gestione di dataset eterogenei.

Hategan et al. (2025) hanno dimostrato l’efficacia dell’integrazione NMR–ML nella classificazione automatica di vini in base a varietà, provenienza geografica e annata. I dati NMR di 63 campioni sono stati elaborati mediante Regressione Logistica e k-Nearest Neighbors, con performance eccellenti. La Regressione Logistica, in particolare, ha raggiunto accuratezze prossime al 100%, confermando il potenziale del ML per l’autenticazione e la tracciabilità dei vini anche su dataset di dimensioni moderate.

Infine, Meoni et al. (2025) hanno applicato modelli Random Forest di regressione e classificazione a dati spettroscopici NMR, con l’obiettivo di predire parametri chimici e sensoriali dell’olio extravergine di oliva, nonché di discriminare l’origine varietale e l’annata di produzione. L’analisi ha coinvolto un’ampia collezione di campioni di EVOO provenienti

da cultivar tipiche italiane, raccolti nel corso di sei annate consecutive. I modelli di regressione RF hanno permesso di stimare con elevata accuratezza numerosi parametri qualitativi dell'olio, tra cui la composizione in acidi grassi, polifenoli, tocoferoli, l'indice di perossidi e diversi attributi sensoriali. Parallelamente, i modelli di classificazione RF hanno consentito l'identificazione dell'origine botanica e dell'anno di raccolta, raggiungendo accuratezze fino al 97,3%. Lo studio ha evidenziato la capacità dell'approccio NMR–ML non solo di discriminare tra varietà di EVOO, ma anche di cogliere variazioni sottili nel profilo metabolico legate all'annata produttiva, aprendo nuove prospettive per la tracciabilità temporale e la valorizzazione delle produzioni tipiche.

Complessivamente, gli studi citati dimostrano come l'integrazione tra spettroscopia NMR e algoritmi di machine learning costituisca una strategia avanzata e promettente per affrontare in modo efficace le sfide legate all'autenticazione, al controllo qualità e alla valorizzazione dei prodotti agroalimentari ad alto valore aggiunto.

2.5 Blockchain e Tracciabilità Smart dei Prodotti Agroalimentari

Negli ultimi anni, la tecnologia blockchain ha suscitato un crescente interesse in una vasta gamma di settori, imponendosi come un paradigma emergente per la gestione sicura, trasparente e decentralizzata delle informazioni digitali. Nata inizialmente come infrastruttura a supporto delle criptovalute, questa tecnologia si è rapidamente evoluta in una piattaforma versatile, capace di abilitare applicazioni complesse in ambiti quali la finanza, la sanità, la logistica, la pubblica amministrazione e, più recentemente, il settore agroalimentare (Casino et al., 2019). Il principio fondante della blockchain risiede nella creazione di un registro distribuito e immutabile, in cui ogni transazione è verificata da una rete di nodi e protetta tramite tecniche crittografiche avanzate. Tale struttura rende pressoché impossibile la modifica retroattiva dei dati senza il consenso collettivo della rete, garantendo così integrità, tracciabilità e fiducia tra i partecipanti (Nakamoto, 2008; Zheng et al., 2018).

Parallelamente, la crescente globalizzazione e la complessità delle filiere agroalimentari hanno evidenziato importanti criticità nella capacità dei sistemi tradizionali di garantire l'autenticità, la sicurezza e la trasparenza dei prodotti lungo l'intero percorso "dal campo alla tavola". I meccanismi di tracciabilità convenzionali, spesso basati su documentazione cartacea o su database centralizzati accessibili solo a determinati attori, mostrano limiti significativi in termini di affidabilità, tempestività e verificabilità delle informazioni. Problemi quali le frodi alimentari, le contraffazioni e la perdita di fiducia da parte dei consumatori nei confronti delle etichette sono diventati questioni sempre più rilevanti a livello globale (Galvez et al., 2018).

In questo contesto, la blockchain si configura come una soluzione tecnologica innovativa, in grado di rivoluzionare le modalità di tracciabilità nel settore agroalimentare. Grazie alla sua natura distribuita, alla trasparenza condivisa e alla protezione crittografica dei dati, essa consente di costruire un registro digitale sicuro, consultabile e verificabile, capace di documentare in modo cronologico e inalterabile tutte le fasi della filiera produttiva: dalla coltivazione delle materie prime, alla loro trasformazione, confezionamento e distribuzione, fino al consumatore finale (Galvez et al., 2018; Tripoli & Schmidhuber, 2018).

Un'applicazione particolarmente significativa di questi principi riguarda la gestione e la certificazione della biomassa legnosa, un materiale caratterizzato da un'elevata eterogeneità in termini di origine, composizione e qualità. In tale ambito, diventa sempre più necessario adottare strumenti che consentano una tracciabilità oggettiva e sicura lungo l'intero ciclo di vita del materiale, dalla raccolta alla caratterizzazione analitica. L'integrazione tra tecniche di

analisi avanzate e tecnologie digitali di registrazione dei dati, come i sensori intelligenti e la blockchain, rappresenta un'evoluzione significativa nei protocolli di certificazione della biomassa. Questo approccio sinergico costituisce la base dell'integrazione della blockchain con tecnologie analitiche avanzate, come approfondito nella sezione seguente.

2.5.1 Integrazione con Tecnologie Analitiche Avanzate

L'adozione di un approccio "smart" alla tracciabilità agroalimentare implica l'integrazione della tecnologia blockchain con strumenti avanzati dell'Industria 4.0, al fine di raccogliere, verificare e validare in modo trasparente i dati di qualità lungo l'intera filiera. In particolare, l'uso combinato di dispositivi IoT, sensori RFID e tecniche analitiche come la spettroscopia NMR, consente di associare a ciascun lotto di prodotto un'impronta digitale univoca, certificando autenticità, origine e conformità ai requisiti di qualità.

2.5.1.1 Sensoristica IoT e RFID

Dispositivi basati su tecnologie IoT (Internet of Things), come sensori ambientali intelligenti, e RFID (Radio Frequency Identification), come tag elettronici applicati ai prodotti, consentono il monitoraggio in tempo reale di parametri critici durante le fasi di coltivazione, trasformazione e distribuzione. I sensori IoT rilevano automaticamente dati quali temperatura, umidità, localizzazione GPS e condizioni logistiche, mentre i tag RFID permettono l'identificazione e il tracciamento univoco dei lotti, anche senza contatto diretto.

Tutte queste informazioni possono essere trasmesse in modo continuo e automatico a una rete blockchain, dove ogni lettura viene registrata in forma immutabile. Gli attori autorizzati della filiera possono così accedere a un registro digitale sicuro e aggiornato, migliorando la trasparenza e la capacità di risposta in caso di anomalie o non conformità (Tian, 2017; Kaur et al., 2022).

2.5.1.2 Integrazione con Spettroscopia NMR

La spettroscopia NMR (Risonanza Magnetica Nucleare) rappresenta una delle tecniche analitiche più avanzate e affidabili per la caratterizzazione chimica degli alimenti. Grazie alla sua capacità di fornire un profilo metabolico dettagliato e riproducibile, l'NMR è ampiamente utilizzata per identificare la composizione molecolare, individuare potenziali adulterazioni, verificare l'origine geografica e la varietà botanica dei prodotti agroalimentari.

L'integrazione dei dati NMR con la tecnologia blockchain apre nuove prospettive per la tracciabilità e la certificazione scientifica della qualità. In questo contesto, i risultati analitici – ad esempio il fingerprint metabolomico caratteristico di un determinato alimento – possono

essere trasformati in hash crittografici e registrati su una blockchain. Questo processo consente di associare a ciascun lotto di prodotto un certificato digitale sicuro, contenente informazioni derivanti direttamente dall'analisi strumentale.

La registrazione in blockchain garantisce l'immutabilità e la verificabilità di tali informazioni: qualunque tentativo di alterazione del dato analitico, o di sostituzione del prodotto fisico, produrrebbe un disallineamento tra il dato atteso e l'hash registrato, rendendo immediatamente evidente la non conformità. In tal modo, la spettroscopia NMR non solo fornisce una base scientifica solida per la valutazione qualitativa dei prodotti, ma diventa anche uno strumento centrale in un sistema digitale di tracciabilità distribuita, trasparente e sicuro (Antonicelli et al., 2021; Lukacs et al., 2025).

2.5.2 Aspetti Tecnici: Funzionamento, Smart Contract, Modelli Pubblici vs Privati e Interoperabilità – Applicazione al contesto agroalimentare

2.5.2.1 Funzionamento della blockchain

Nel contesto agroalimentare, una rete blockchain viene solitamente configurata come un'infrastruttura condivisa alla quale partecipano attivamente tutti gli attori della filiera: produttori agricoli, trasformatori, distributori, rivenditori e organismi di certificazione. In base al tipo di architettura scelta, la rete può essere pubblica, cioè aperta a qualsiasi partecipante (modello meno comune nel settore alimentare per ragioni di privacy e governance), oppure privata o consortile, quindi riservata a membri autorizzati e già noti tra loro. Quest'ultima opzione è la più adottata nel settore, poiché consente di bilanciare trasparenza e riservatezza, garantendo al contempo un controllo più efficace sui partecipanti e sulle informazioni condivise.

Ogni qualvolta un attore della filiera esegue un'operazione rilevante su un lotto – come la raccolta della materia prima, un'analisi di laboratorio o il confezionamento del prodotto – viene generata una transazione digitale che descrive nel dettaglio l'evento. Un esempio può essere la registrazione del raccolto di un lotto di mele con data, luogo, quantità e nome dell'azienda agricola. Questa informazione viene trasmessa alla rete, validata da uno o più nodi secondo il protocollo di consenso adottato, e infine inserita in un blocco. I protocolli di consenso variano: nelle blockchain pubbliche si utilizzano spesso algoritmi come Proof of Work o Proof of Stake, mentre nelle reti private o consortili si preferiscono meccanismi più efficienti, come quelli

basati su Byzantine Fault Tolerance, che garantiscono una maggiore velocità e sicurezza nelle transazioni (Zheng et al., 2018).

Una volta validato, il blocco viene agganciato alla catena esistente grazie a un collegamento crittografico (l'hash del blocco precedente), e viene replicato su tutti i nodi della rete. Questo processo rende la blockchain un registro cronologico e immutabile: qualsiasi tentativo di alterazione dei dati genererebbe una discrepanza tra i blocchi e richiederebbe un nuovo consenso della rete, operazione tecnicamente ed organizzativamente impraticabile (Casino et al., 2019).

Nel settore agroalimentare, la blockchain assume così il ruolo di un libro mastro digitale condiviso, in cui viene registrata tutta la "storia" di un prodotto. Per esempio, un lotto di latte può essere tracciato attraverso una sequenza di blocchi che documentano il rilascio del certificato veterinario, la raccolta del latte, l'ingresso nel caseificio, le condizioni ambientali, i risultati delle analisi di qualità, il confezionamento e la distribuzione finale. Ogni blocco è associato a un ID lotto, permettendo di ricostruire l'intera filiera in modo trasparente e verificabile. A supporto della fruibilità, le piattaforme blockchain più evolute forniscono interfacce grafiche e strumenti di visualizzazione che permettono di accedere in modo selettivo alle informazioni pertinenti, sia per gli operatori della filiera che per consumatori o autorità di controllo (Galvez et al., 2018).

2.5.2.2 Automazione con smart contract

Gli smart contract, ovvero programmi auto-eseguibili residenti sulla blockchain, permettono di automatizzare processi lungo la filiera in base a condizioni prestabilite. Quando un determinato evento si verifica – come il superamento di una soglia critica di temperatura, o il caricamento di un certificato analitico – lo smart contract si attiva e può generare un allarme, inviare notifiche o registrare un nuovo evento sulla catena. Nel settore agroalimentare, ciò consente un controllo qualità dinamico, automatico e trasparente, riducendo la necessità di interventi manuali e migliorando l'affidabilità complessiva del sistema (Kamilaris et al., 2019; Koutras et al., 2023).

L'utilizzo di questi contratti introduce una logica computazionale distribuita, nella quale la blockchain non si limita più a memorizzare dati passivamente, ma è in grado di applicare regole operative in modo automatico (Zheng et al., 2018). Tuttavia, la programmazione degli smart contract richiede grande attenzione: un errore nel codice può produrre effetti indesiderati e irreversibili (Atzei et al., 2017). Inoltre, è essenziale garantire l'integrità dei dati provenienti dall'esterno – come quelli generati da sensori o sistemi di laboratorio – che fungono da input

per l'attivazione dei contratti. Per questo motivo, nelle implementazioni consortili, è prassi che gli smart contract vengano sviluppati da fornitori qualificati e approvati collegialmente dai membri della rete, prima del loro utilizzo operativo (Casino et al., 2019).

2.5.2.3 Blockchain pubbliche e private: caratteristiche e implicazioni

La scelta tra blockchain pubblica e privata rappresenta una delle decisioni strategiche principali in fase di progettazione. Le blockchain pubbliche – come Ethereum, Algorand o Tezos – sono decentralizzate per definizione: nessun ente centrale le controlla, chiunque può partecipare e verificare le transazioni. Questo modello assicura un elevato livello di trasparenza e fiducia "super partes", grazie alla presenza di migliaia di nodi distribuiti. Tuttavia, queste reti mostrano limiti in termini di performance, hanno costi variabili legati all'uso di criptovalute e non offrono garanzie adeguate in termini di riservatezza, a meno di utilizzare meccanismi crittografici avanzati.

Le blockchain private o consortili, come Hyperledger Fabric, R3 Corda o Quorum, sono invece progettate per operare in contesti chiusi. In questi sistemi, solo gli attori previamente autorizzati possono accedere alla rete, leggere o scrivere informazioni. Questo approccio consente una gestione più efficiente, l'assenza di costi transazionali legati alle criptovalute e una maggiore adattabilità alle policy aziendali. Alcune piattaforme, come Fabric, permettono anche la creazione di canali di comunicazione privati, visibili solo a un sottoinsieme di partecipanti, il che è particolarmente utile in caso di dati sensibili.

Il principale limite delle reti permissioned risiede nella minore decentralizzazione. La governance è affidata a un gruppo ristretto di attori, che può modificarne le regole operative o l'accesso ai nodi, il che richiede un livello iniziale di fiducia tra i partecipanti. Nel settore agroalimentare, questo compromesso risulta comunque accettabile, motivo per cui la maggior parte dei progetti reali adotta proprio reti consortili. Un esempio significativo è rappresentato dal progetto ungherese per la tracciabilità della patata dolce, basato su Hyperledger e coordinato da un consorzio di filiera (Lukacs et al., 2025).

Una soluzione ibrida particolarmente promettente consiste nel gestire i dati operativi all'interno di una rete permissioned, pubblicando periodicamente l'hash crittografico dei blocchi su una blockchain pubblica. In tal modo, anche se i membri del consorzio modificassero i dati interni, la discrepanza con gli hash pubblici renderebbe immediatamente evidente qualsiasi tentativo di manomissione (Herbke et al., 2024).

2.5.2.4 Interoperabilità e integrazione con i sistemi informativi

Uno dei principali obiettivi tecnici nella progettazione di un sistema blockchain per la tracciabilità alimentare è l'interoperabilità. I dati dovrebbero poter fluire in modo trasparente tra piattaforme diverse, evitando blocchi informativi o dipendenze da fornitori specifici. Questo è possibile solo attraverso la standardizzazione dei dati e l'adozione di formati condivisi a livello internazionale. L'iniziativa GS1 US, ad esempio, ha definito linee guida per l'utilizzo di identificatori globali (come il GTIN) e messaggi EPCIS all'interno di reti blockchain, dimostrando l'efficacia dell'approccio standardizzato per garantire la compatibilità tra sistemi (GS1 US., 2020).

Dal punto di vista tecnico, un sistema blockchain non può sostituire i software gestionali aziendali, ma deve integrarsi con essi. Questo vale per i sistemi gestionali esistenti (ERP, MES, database di magazzino), così come per i dispositivi fisici come sensori, etichettatrici e lettori RFID. Per assicurare una comunicazione fluida, è necessario sviluppare interfacce applicative (API) robuste e middleware in grado di convertire gli eventi aziendali in transazioni blockchain. Inoltre, l'utilizzo di gateway IoT e servizi cloud permette di raccogliere i dati grezzi dai dispositivi, elaborarli e trasmettere alla blockchain solo le informazioni essenziali, riducendo il carico della rete e garantendo maggiore stabilità (Kamilaris et al., 2019).

2.5.2.5 Sicurezza, gestione delle chiavi e continuità operativa

Anche se la blockchain garantisce l'immutabilità e l'integrità dei dati, la sicurezza dell'infrastruttura complessiva dipende da altri elementi critici. I nodi della rete devono essere protetti contro accessi non autorizzati e mantenuti operativi in modo sicuro. Un aspetto particolarmente delicato riguarda la gestione delle chiavi private: la loro perdita rende impossibile per l'utente firmare transazioni, mentre il loro furto può compromettere l'affidabilità dell'intero sistema. È quindi necessario adottare soluzioni professionali per il key management, come sistemi di backup crittografati o l'utilizzo di dispositivi HSM (Hardware Security Module).

Infine, è importante pianificare la continuità operativa. In caso di guasti della rete o interruzioni temporanee, la filiera deve poter continuare a funzionare, registrando i dati in locale per poi sincronizzarli con la blockchain una volta ristabilita la connessione. Le reti consortili ben progettate, basate su infrastrutture ridondanti distribuite su cloud e on-premise, sono in grado di offrire livelli elevati di affidabilità e disponibilità (Bitquery, 2024).

2.5.3 Vantaggi della Registrazione in Blockchain rispetto ai Sistemi Tradizionali

L'adozione della tecnologia blockchain nella tracciabilità agroalimentare offre numerosi vantaggi rispetto ai sistemi tradizionali basati su documentazione cartacea o database centralizzati (Galvez et al., 2018; Casino et al., 2019).

Uno dei principali punti di forza della blockchain risiede nella immutabilità e integrità dei dati. Una volta registrata, ogni informazione diventa permanente: non può essere modificata né cancellata senza il consenso della rete. L'architettura crittografica della blockchain, in cui ciascun blocco contiene l'hash del blocco precedente, rende ogni alterazione immediatamente rilevabile, prevenendo modifiche fraudolente o involontarie. Questo garantisce un livello di sicurezza e affidabilità difficilmente ottenibile nei sistemi centralizzati.

La trasparenza e visibilità condivisa della blockchain è un ulteriore elemento distintivo. In una rete distribuita – pubblica o consortile – tutte le parti autorizzate possono accedere alle informazioni di filiera in tempo reale. Questo migliora il controllo da parte delle autorità, aumenta la responsabilizzazione degli operatori e rafforza la fiducia del consumatore. Ogni azione documentata – dalla raccolta alla distribuzione – è tracciabile e consultabile, favorendo una cultura della qualità e della responsabilità.

La blockchain introduce inoltre un nuovo paradigma di fiducia decentralizzata. Nei sistemi tradizionali, la fiducia si basa su enti centrali o sulla reputazione dei singoli attori. Con la blockchain, invece, la fiducia è incorporata nell'infrastruttura stessa: ogni transazione è verificata dalla rete e garantita crittograficamente. Questo consente a soggetti anche non interconnessi tra loro di collaborare in modo sicuro e trasparente, un aspetto particolarmente rilevante in filiere complesse e globali.

Tra i benefici più concreti si annovera anche la riduzione delle frodi alimentari e il rafforzamento della sicurezza. Poiché ogni fase della filiera lascia una traccia immutabile, eventuali anomalie – come l'assenza di certificazioni obbligatorie o parametri chimici incoerenti – possono essere facilmente identificate. Questo rende più difficile perpetrare pratiche fraudolente e facilita l'individuazione precisa del punto critico all'interno della catena.

Un altro vantaggio rilevante è l'aumento dell'efficienza operativa. La disponibilità di una catena informativa accessibile e interrogabile in tempo reale consente di abbreviare drasticamente i tempi per la verifica dei lotti, l'esecuzione degli audit e la gestione delle emergenze. In caso di allerte sanitarie, ad esempio, la blockchain permette di rintracciare rapidamente i lotti coinvolti e intervenire in modo mirato, evitando richiami generalizzati e

costosi. La possibilità di disporre di un *digital twin* – un gemello digitale per ogni lotto contenente tutte le informazioni registrate in blockchain – rende più semplice il controllo documentale, automatizza le verifiche e riduce l'onere amministrativo.

Infine, l'integrazione con smart contract permette di automatizzare processi lungo la catena del valore, come l'esecuzione di pagamenti, la gestione di penali per ritardi o le verifiche di conformità, riducendo i tempi di elaborazione e migliorando l'efficienza complessiva del sistema.

2.5.4 Limiti e Criticità della Blockchain in Filiera

Nonostante il potenziale innovativo della blockchain, la sua applicazione nelle filiere agroalimentari presenta ancora alcune sfide tecniche, economiche e organizzative che ne limitano la diffusione su larga scala.

Scalabilità e performance: Le blockchain pubbliche di prima generazione, come Bitcoin o Ethereum, soffrono di limiti strutturali in termini di velocità e costo delle transazioni, rendendole poco adatte a contesti caratterizzati da un'elevata frequenza di eventi, come le filiere alimentari. Anche le blockchain permissioned, adottate nei consorzi privati, richiedono infrastrutture robuste per gestire in modo efficiente grandi volumi di dati. Per mitigare questi limiti, sono stati proposti sistemi più performanti basati su meccanismi di consenso alternativi (es. Proof-of-Stake) e l'uso di strategie di compressione dei dati, come la registrazione in blockchain del solo hash crittografico o l'adozione di architetture multilivello (side-chain, off-chain) (Galvez et al., 2018).

Costi e accessibilità tecnologica: L'implementazione della blockchain implica costi rilevanti legati allo sviluppo software, all'integrazione con sistemi preesistenti, alla gestione di nodi e alla sicurezza delle chiavi digitali. L'adozione di piattaforme enterprise (come Hyperledger o IBM Food Trust) può risultare onerosa, soprattutto per le PMI del settore agroalimentare, che spesso operano con margini economici ristretti. Inoltre, la carenza di infrastrutture digitali e competenze specifiche in alcune aree rurali rappresenta una barriera concreta all'utilizzo diffuso della tecnologia (Kamilaris et al., 2019).

Adozione eterogenea e resistenza al cambiamento: Il successo di un sistema di tracciabilità distribuita richiede il coinvolgimento attivo di tutti gli attori della filiera. Tuttavia, la variabilità nei livelli di digitalizzazione e la mancanza di consapevolezza tecnica possono generare resistenze, in particolare tra i piccoli produttori o fornitori marginali (Kamilaris et al., 2019). A queste difficoltà si sommano preoccupazioni legate alla condivisione di informazioni sensibili, che talvolta vengono percepite come una minaccia al vantaggio competitivo. In questo contesto,

la disponibilità di modelli di accesso selettivo e incentivi economici (es. accesso a nuovi mercati o certificazioni premium) si rivela cruciale per favorire l'adozione.

Interoperabilità e standardizzazione dei dati: L'assenza di standard condivisi nella raccolta e nella formattazione dei dati costituisce un serio ostacolo all'integrazione tra attori e piattaforme. Differenze nei codici di lotto, nelle unità di misura o nei metadati possono compromettere la coerenza informativa e ridurre l'affidabilità del sistema. La blockchain, per sua natura decentralizzata, richiede uno sforzo collettivo di armonizzazione tra consorzi, enti normativi e sviluppatori, affinché i dati tracciati siano interoperabili e confrontabili su scala internazionale (Francisco & Swanson, 2018).

Privacy e gestione dei dati sensibili: L'elevata trasparenza della blockchain può risultare problematica in presenza di informazioni commerciali o personali. Nelle implementazioni consortili, il controllo granulare degli accessi consente di proteggere i dati più sensibili, ma è fondamentale adottare architetture ibride che prevedano la registrazione sulla blockchain del solo hash crittografico, mantenendo i dati completi off-chain (Lukacs et al., 2025). Questo approccio consente di garantire l'integrità delle informazioni senza comprometterne la riservatezza. Inoltre, è necessario che ogni sistema sia progettato nel rispetto delle normative vigenti in materia di protezione dei dati personali, come il Regolamento GDPR.

Connessione tra mondo fisico e digitale: Un altro limite strutturale della blockchain è la sua incapacità di verificare la veridicità dei dati al momento dell'immissione. In assenza di meccanismi di validazione esterna, il sistema rimane vulnerabile a informazioni false o parziali ("garbage in, garbage out"). Per superare questa criticità, è essenziale integrare la blockchain con tecnologie affidabili per la raccolta dati, come sensori IoT certificati, marcatori chimici o analisi metabolomiche eseguite da enti accreditati. L'uso di firme digitali non ripudiabili da parte di operatori certificati consente inoltre di associare ogni informazione a un responsabile identificabile, rafforzando la tracciabilità e la responsabilità legale (Kamilaris et al., 2019; Lukacs et al., 2025).

2.5.5 Esempi Pratici e Casi di Studio

Numerosi studi in letteratura evidenziano le potenzialità della tecnologia blockchain nel settore agroalimentare, in particolare per supportare la tracciabilità dei prodotti, la certificazione di origine, la garanzia della qualità e la sostenibilità lungo le filiere produttive. Tuttavia, i casi concreti di applicazione reale risultano ancora limitati, spesso circoscritti a progetti pilota o a iniziative sperimentali in fase di sviluppo (Antonucci et al., 2019; Casino et al., 2019; Ellahi et

al., 2024). Nel seguito si riportano alcuni casi studio significativi che dimostrano l'integrazione della blockchain nelle filiere agroalimentari.

Blockchain per la tracciabilità e la gestione efficiente della filiera delle uova: un caso studio statunitense

Uno degli esempi riguarda il settore della distribuzione delle uova negli Stati Uniti, dove è stata implementata una soluzione blockchain per tracciare ogni fase – dalla raccolta in azienda agricola fino allo scaffale del supermercato – utilizzando sensori IoT e smart contract per registrare parametri critici come temperatura e umidità in tempo reale (Bumblauskas et al., 2020).

Nel progetto descritto, denominato Bytable, è stata adottata una blockchain permissioned (Hyperledger Sawtooth) per registrare i dati generati da circa 100 fattorie locali. Le informazioni associate a ogni lotto di uova – incluso l'orario di raccolta, le condizioni ambientali, i dati di confezionamento e distribuzione – venivano salvate in modo immutabile nel ledger, rendendole accessibili tramite QR code direttamente sulla confezione. I consumatori potevano così accedere a una piattaforma dedicata per consultare il profilo completo del prodotto acquistato, incluso il rispetto dei requisiti normativi e gli eventuali certificati di benessere animale.

I risultati del proof-of-concept sono stati positivi: l'adozione da parte dei consumatori ha superato le aspettative, con tassi di interazione superiori al 20%, dimostrando un interesse concreto verso la trasparenza alimentare. Dal punto di vista tecnico, l'architettura del sistema ha permesso di integrare i dati raccolti da sensori distribuiti lungo la catena, migliorando la qualità dei dati e riducendo il rischio di errore umano.

Tracciabilità blockchain e sostenibilità ambientale nella produzione di Fontina DOP: un modello replicabile

nel settore lattiero-caseario, un caso studio emblematico è rappresentato dal progetto europeo *TYPICALP*, che ha visto l'implementazione di una piattaforma di tracciabilità basata su *green blockchain* per la filiera della Fontina DOP prodotta in Valle d'Aosta.

Il sistema proposto da Varavallo et al. (2022) si basa sulla blockchain Algorand, una soluzione che impiega un meccanismo di consenso Pure Proof-of-Stake (PPoS), caratterizzato da basso consumo energetico, alta scalabilità e ridotti costi per transazione. Tutti gli attori della filiera – dagli allevatori ai caseifici, dai trasportatori ai consorzi di controllo – partecipano attivamente alla registrazione degli eventi critici, tra cui mungitura, trasformazione, stagionatura e confezionamento. I dati vengono salvati in un sistema gestionale e successivamente registrati

in modo immutabile sulla blockchain durante la fase di packaging, garantendo così l'autenticità e la non alterabilità delle informazioni.

Un elemento distintivo di questo approccio è la generazione automatica di un QR code per ogni lotto di formaggio porzionato. Il consumatore può scansionarlo per accedere a una pagina web che riporta tutte le informazioni verificate lungo la filiera, dal latte utilizzato fino alla data di confezionamento. Tale sistema promuove trasparenza, fiducia e valorizzazione del prodotto locale, riducendo i rischi legati a frodi o falsificazioni e facilitando al contempo i controlli da parte delle autorità competenti.

L'applicazione di blockchain in questo contesto non si limita alla sicurezza alimentare: favorisce anche la sostenibilità ambientale e l'efficienza gestionale.

Blockchain per la gestione del commercio agroalimentare: il caso FTSCON in Cina

Oltre alla tracciabilità e all'autenticazione dei prodotti, la blockchain trova un'applicazione strategica anche nella gestione delle transazioni commerciali lungo le filiere agroalimentari, soprattutto nei contesti caratterizzati da alti livelli di frammentazione e scarsa fiducia tra gli attori.

Un esempio di particolare rilievo è rappresentato dal sistema FTSCON (*Food Trading System with Consortium Blockchain*), descritto da Mao et al. (2018), progettato per supportare la gestione sostenibile, trasparente e credibile del mercato alimentare nella provincia cinese dello Shandong. Il sistema nasce per risolvere alcune criticità ricorrenti nel commercio locale di prodotti agricoli, come la mancanza di standardizzazione, la bassa visibilità delle transazioni e la disparità di potere contrattuale tra piccoli produttori e grossisti.

Il sistema FTSCON si basa su una blockchain consortile che coinvolge diversi attori della filiera: produttori, distributori, regolatori pubblici e fornitori di servizi logistici. Attraverso l'uso combinato di smart contract e funzioni di matching automatico, la piattaforma consente di registrare ogni fase della negoziazione commerciale in modo sicuro, trasparente e immutabile. Le transazioni vengono validate collettivamente dai nodi autorizzati, riducendo il rischio di manipolazioni o discrepanze nei dati. La blockchain non è impiegata solo per garantire la fiducia tra le parti, ma anche per ottimizzare l'efficienza del mercato attraverso un algoritmo di abbinamento intelligente tra domanda e offerta.

Lo studio evidenzia come questo approccio abbia favorito una maggiore trasparenza delle informazioni sul prodotto e sui prezzi, accessibili in tempo reale a tutti gli attori. Inoltre, ha permesso di ridurre significativamente i tempi di contrattazione e i costi operativi, migliorando la redditività per i piccoli produttori agricoli e contribuendo a una maggiore stabilità del

mercato. FTSCON si è dimostrato uno strumento efficace nel rafforzare l'equità commerciale, riducendo l'intermediazione e facilitando relazioni più dirette e tracciabili tra i soggetti della filiera.

Dal punto di vista tecnologico, il sistema è stato progettato per integrarsi con i gestionali aziendali e con dispositivi IoT, consentendo l'acquisizione automatica di dati ambientali e logistici – come temperatura, umidità e geolocalizzazione – che vengono poi associati alle transazioni blockchain. Questo ha rafforzato ulteriormente l'affidabilità e la ricchezza informativa del sistema.

In sintesi, il caso FTSCON dimostra come la blockchain possa essere adottata con successo non solo per tracciare la provenienza dei prodotti, ma anche come infrastruttura tecnologica capace di migliorare l'efficienza, la trasparenza e l'integrità delle dinamiche di mercato nelle filiere agroalimentari.

Capitolo 3 Materiali e Metodi

Questo capitolo descrive le tecniche adottate per l'analisi di diverse matrici agroalimentari, comprendenti esperimenti NMR, procedure di elaborazione dei dati e l'applicazione integrata di metodi statistici e algoritmi di machine learning per la classificazione dei campioni. L'obiettivo è caratterizzare i profili metabolici e contribuire allo sviluppo di protocolli a supporto della certificazione di qualità e origine dei prodotti analizzati.
Inoltre, viene presentata un'analisi preliminare sui campioni di cippato di legno, finalizzata a testare l'integrazione dei dati metabolici ottenuti mediante spettroscopia NMR con quelli acquisiti da sistemi di sensori IoT e registrati in blockchain, nell'ottica di sviluppare soluzioni innovative per la tracciabilità e il monitoraggio automatico dei materiali lungo la filiera.

3.1 Gestione delle matrici sperimentali

Per assicurare un'indagine rappresentativa e scientificamente solida, sono state selezionate matrici di particolare rilevanza nel contesto agroalimentare: il latte UHT, il frumento e il concentrato di pomodoro. Si tratta di prodotti di largo consumo, caratterizzati da un impatto economico significativo sia a livello nazionale che internazionale. La scelta di queste matrici è stata guidata dalla volontà di sviluppare protocolli analitici che potessero rispondere alle esigenze di tracciabilità e autenticazione lungo le filiere produttive, contribuendo così alla valorizzazione della qualità e della trasparenza, in coerenza con i crescenti standard richiesti dal mercato e dai consumatori.
Lo studio è stato inoltre esteso verso applicazioni agro-forestali, includendo tra le matrici analizzate anche i cippati di legno, con l'obiettivo di esplorare le caratteristiche chimiche di materiali vegetali a diverso potenziale applicativo e già oggetto di interesse in sistemi avanzati di tracciabilità. L'interesse per la caratterizzazione dei cippati si inserisce inoltre in un più ampio sforzo di valorizzazione della biomassa nel quadro dell'economia circolare e della transizione ecologica, con implicazioni rilevanti nei settori dell'energia, dell'agronomia e della chimica verde (Vuković & Tišma, 2024).
La raccolta dei campioni è stata possibile grazie alla collaborazione con partner strategici. L'azienda Farzati S.p.A. ha fornito supporto operativo e tecnologico, mettendo a disposizione campioni di frumento appartenenti a diverse varietà, oltre a quelli di cippati. Il centro di ricerca CREA di Foggia ha contribuito all'analisi fornendo campioni della varietà "Senatore Cappelli", mentre il professor Romano dell'Università degli Studi di Napoli Federico II ha reso disponibili i campioni di latte UHT e di concentrato di pomodoro. Tutti i campioni sono stati raccolti e

preparati seguendo protocolli rigorosi, calibrati in funzione della natura della matrice, con l'obiettivo di preservare l'integrità dei metaboliti e garantire l'affidabilità delle analisi successive.

Le liste complete delle matrici analizzate mediante spettroscopia HR-MAS NMR e CP-MAS NMR sono riportate in Tabella 3.1 e Tabella 3.2. I dettagli relativi alla preparazione dei campioni sono descritti nella sezione successiva.

Tabella 3.1 – Elenco delle matrici analizzate mediante spettroscopia HR-MAS NMR

Matrice	**Classi (Provenienza)**
Frumento Farzati	Achille (T. durum - San Severino Marche (MC))
	Creso (T. durum - San Severino Marche (MC))
	Minosse (T. durum - San Severino Marche (MC))
	Telemaco (T. durum - San Severino Marche (MC))
	Norberto (T. monococcum - San Severino Marche (MC))
	Hammurabi (T. monococcum - San Severino Marche (MC))
Frumento var. Senatore Cappelli	Cavaliere (Troia (FG)-Puglia)
	Ciacci (Urbino (PU)-Marche)
	Colucci (Bovino (FG)-Puglia)
	Crea1C (Troia (FG)-Puglia)
	De Ioanni (Benevento (BN) -Campania)
Pomodoro Concentrato	Italia
	Cina
	Spagna
Latte UHT	Arborea Intero (Italiano)
	Arborea Parzialmente Scremato (Italiano)
	Arborea Scremato (Italiano)
	Berna Parzialmente Scremato (Europeo)
	Berna Scremato (Europeo)
	Berna Intero (Italiano)
	Berna Parzialmente Scremato (Italiano)
	Granarolo Parzialmente Scremato (Italiano)
	Matese Intero (Europeo)
	Matese Parzialmente Scremato (Europeo)
	Matese Scremato (Europeo)

Tabella 3.2 – Matrice (cippati di legno) analizzata mediante spettroscopia 13C CP-MAS NMR con indicazione di specie, altezza e zona di prelievo.

Cippati di legno		
Specie Arborea	**Altitudine di Prelievo (m)**	**Zona di Prelievo**
Abete Rosso	1481	Trentino-Alto Adige
Abete Rosso	328	Lombardia
Castagno	173	Toscana
Castagno	252	Molise
Cerro	197	Umbria
Cerro	243	Emilia-Romagna
Cerro	321	Calabria
Douglasia	238	Calabria
Douglasia	342	Basilicata
Faggio	141	Campania
Faggio	340	Calabria
Faggio	346	Calabria
Pino Laricio	1383	Calabria
Pino Laricio	217	Puglia
Pino Nero	1348	Toscana
Pino Nero	1350	Trentino-Alto Adige
Pino Nero	255	Calabria
Pioppo	139	Calabria
Pioppo	145	Lombardia
Pioppo	327	Umbria

3.1.1 Preparazione dei campioni

3.1.1.1 Frumento

Il frumento è stato selezionato come matrice di studio per la sua rilevanza economica e diffusione a livello nazionale e internazionale, nonché per l'interesse crescente verso la certificazione di origine e autenticità nelle filiere cerealicole. Sono state condotte due linee di ricerca: la prima si è focalizzata sulle differenze tra varietà di frumento moderno e antico, confrontando nel dettaglio campioni appartenenti a *Triticum durum* (frumento duro moderno: Achille, Creso, Telemaco e Minosse) e a *Triticum monococcum* (farro monococco antico: Norberto e Hammurabi), mentre la seconda ha analizzato i profili metabolici all'interno della varietà "Senatore Cappelli", coltivata in diverse aree geografiche. Il prelievo in campo è stato condotto seguendo un modello a griglia, con punti di prelievo distribuiti uniformemente

all'interno di ciascuna parcella in triplicato. I semi di frumento sono stati sottoposti a macinazione mediante mulino a pietra "Mockmill 200", scelto per la sua capacità di preservare l'integrità dei metaboliti attraverso una frantumazione delicata e uniforme. Durante il processo, è stato attentamente monitorato l'innalzamento termico delle pietre, al fine di evitare alterazioni nella composizione chimica del campione. Le farine così ottenute, omogenee e rappresentative, sono state quindi stoccate e successivamente poste in essiccatore per garantirne la stabilità nel tempo.

3.1.1.2 Concentrato di pomodoro

Sono stati analizzati campioni di concentrato di pomodoro provenienti da tre distinte aree geografiche – Italia, Cina e Spagna – con l'obiettivo di confrontarne i profili metabolici in funzione dell'origine e di individuare eventuali marcatori distintivi associabili alla zona di produzione e alle specifiche tecniche di trasformazione. I campioni, trasferiti in provette Falcon da 50 mL, sono stati sottoposti a liofilizzazione per la completa rimozione dell'umidità residua e successivamente conservati in essiccatore, al fine di garantirne la stabilità e preservare l'integrità metabolica.

3.1.1.3 Latte UHT

L'analisi ha riguardato campioni di latte pastorizzato (UHT) provenienti da diverse marche nazionali: Granarolo (GR), Arborea (AR), Berna (BR) e Matese (MT), nelle varianti intero, parzialmente scremato e scremato. L'obiettivo principale era analizzare la variabilità metabolica in funzione della provenienza e del contenuto lipidico, al fine di individuare potenziali marcatori distintivi tra le diverse tipologie. I campioni, raccolti e trasferiti in provette Falcon da 50 mL, sono stati sottoposti a liofilizzazione per garantire la completa rimozione dell'umidità e preservare l'integrità dei metaboliti. Successivamente, sono stati conservati in essiccatore in condizioni controllate, per evitarne l'alterazione dovuta all'assorbimento di umidità residua.

3.1.1.4 Cippati di legno

I *cippati* costituiscono una forma di biomassa lignocellulosica ampiamente utilizzata e oggetto di crescente attenzione sia in ambito applicativo che scientifico. Si tratta di frammenti di materiale legnoso ottenuti attraverso un processo meccanico di triturazione, realizzato per mezzo di cippatrici, che consente di ridurre residui forestali, potature, scarti agricoli e arboricoli in pezzi di dimensioni controllate, solitamente comprese tra 1 e 5 cm.

Ai fini del presente lavoro, i campioni di legno sono stati prelevati da diverse specie arboree, tra cui *Abete rosso, Castagno, Cerro, Douglasia, Faggio, Pino laricio, Pino nero* e *Pioppo*,

selezionate per rappresentare un'ampia varietà di condizioni ecologiche. Il campionamento ha interessato differenti regioni italiane – dal Trentino-Alto Adige alla Calabria – e si è svolto a quote comprese tra 139 e 1481 metri sul livello del mare. Questa strategia ha permesso di includere una gamma rappresentativa di condizioni pedoclimatiche. Ogni campione è stato identificato mediante QR code univoco per garantirne la tracciabilità lungo tutte le fasi analitiche. Il legno è stato poi sottoposto a uno sminuzzamento fine con lame. La polvere così ottenuta è stata trasferita in contenitori idonei e posta in essiccatore a temperatura costante, al fine di minimizzare interferenze dovute all'umidità e assicurare l'affidabilità analitica.

3.1.2 Acquisizione degli spettri HR-MAS e CP-MAS

Le analisi NMR sono state condotte mediante due approcci sperimentali distinti, selezionati in base alla natura della matrice: la tecnica HR-MAS NMR (High-Resolution Magic Angle Spinning) è stata applicata alle matrici alimentari, mentre per i campioni di cippato, in virtù della loro composizione dominata da cellulosa, emicellulose e lignina, è stata utilizzata la tecnica CP-MAS NMR (Cross-Polarization Magic Angle Spinning). Tali tecniche hanno consentito l'analisi non distruttiva di campioni solidi e semi-solidi, garantendo la conservazione delle caratteristiche chimiche e strutturali dei metaboliti. In particolare, l'utilizzo di HR-MAS NMR si è rivelato fondamentale per ottenere spettri ad alta risoluzione, ricchi di informazioni utili per la caratterizzazione dei profili metabolici.

3.1.2.1 Spettroscopia HR-MAS ^{1}H ^{13}C-NMR

Le analisi sono state condotte utilizzando uno spettrometro Bruker Avance 400 MHz (Bruker Biospin, Rheinstetten, Germany), dotato di un campo magnetico da 9.4 T e di una sonda HR-MAS ^{1}H-^{13}C, operante a 400.13 MHz per il ^{1}H e 101.5 MHz per il ^{13}C. I campioni pesati (15 mg ±0.5 per i frumenti, 16mg ±0.5 per il latte e 20mg ± 0.5 per i concentrati di pomodoro) sono stati trasferiti in rotori in zirconia da 4 mm, specifici per l'analisi HR-MAS, muniti di inserti forati in Teflon. Per migliorare la stabilità del sistema ed evitare interferenze spettrali, a ciascun campione è stata aggiunta una soluzione D_2O (99.8% D_2O/H_2O, Armar Chemicals) di circa 35±5 µL contenente lo standard interno 3-(trimetilsilil) propionato-2,2,3,3-d_4 sodico (TMSPA). Il rotore così impaccato è stato sigillato con un tappo in Kel-F. In Figura 3.1 è illustrata la preparazione del campione per l'analisi del frumento, con la polvere liofilizzata ottenuta ed il rotore da 4 mm completo dei suoi componenti.

Figura 3.1 - Preparazione del campione per analisi HR-MAS NMR con evidenza su rotore e sue componenti (inserto forato in Teflon, vite di chiusura e tappo in Kel-F).

L'acquisizione degli spettri è stata effettuata a temperatura controllata (298 ± 1 K). I principali parametri sperimentali ottimizzati comprendevano un tempo di recupero di 2 s, l'applicazione di una sequenza di impulsi *zgpr* con angolo di 90° e tempi di impulso compresi tra 4.5 e 6.0 µs. La larghezza della finestra spettrale è stata fissata a 12 ppm (4789,272 Hz), con una risoluzione spettrale ottenuta acquisendo 32k punti. Il numero di transienti accumulati è stato pari a 256 per i concentrati di pomodoro e a 128 per le altre matrici analizzate. La soppressione del segnale dell'acqua residua è stata ottenuta mediante presaturazione, con un'attenuazione impostata a 60 ± 3 dB.

L'analisi dei metaboliti è stata condotta mediante esperimenti monodimensionali e bidimensionali, ottimizzati per garantire un'elevata qualità spettrale.

L'esperimento ^{1}H–^{1}H COSY è stato condotto utilizzando la sequenza di impulsi *cosygpmfqf*, in modalità Double Quantum Detection (DQD). I dati sono stati acquisiti con 2048 punti (TD) lungo F_2 e 256 incrementi lungo F_1, con un SI impostato a 1024 per F_2 e 256 per F_1, 48 scansioni, 16 dummy scans e tempi di acquisizione (AQ) pari a 0.1597 s (F_2) e 0.0200 s (F_1). La larghezza della finestra spettrale è stata fissata a 16 ppm sia in F_2 sia in F_1.

L'esperimento ^{1}H–^{1}H TOCSY è stato eseguito utilizzando la sequenza di impulsi *mlevgpph19*, con acquisizione di 2048 punti (F_2) e 256 incrementi (F_1), 48 scansioni, 16 dummy scans e larghezza della finestra spettrale impostata a 12 ppm sia in F_2 sia in F_1. Il tempo di mixing per il trasferimento della magnetizzazione è stato impostato a 80 ms.

L’esperimento $^{1}H–^{13}C$ HSQC è stato acquisito utilizzando la sequenza di impulsi *hsqcetgpsisp2.2*, in modalità Echo-Antiecho. I dati sono stati raccolti con 2048 punti (F_2), 256 incrementi (F_1), 80 scansioni e 16 dummy scans. La larghezza della finestra spettrale era di 16 ppm per il ^{1}H e 250 ppm per il ^{13}C.

Questa configurazione sperimentale ha consentito l’acquisizione di spettri ad alta risoluzione, facilitando l’identificazione accurata dei metaboliti presenti nelle matrici analizzate e fornendo una base solida per le successive fasi di elaborazione e modellazione tramite tecniche di intelligenza artificiale.

3.1.2.2 Spettroscopia CP-MAS ^{13}C-NMR

Gli spettri ^{13}C-NMR a polarizzazione incrociata con rotazione ad angolo magico (CP-MAS) sono stati acquisiti utilizzando uno spettrometro Bruker AVANCE™ 300 (operante a 75.475 MHz per il ^{13}C), equipaggiato con una sonda MAS Wide Bore da 4 mm. I campioni solidi (circa 100 mg) sono stati inseriti in rotori di zirconia da 4 mm con tappi in Kel-F e fatti ruotare a una velocità di 10000 ± 1 Hz. L’esperimento è stato condotto applicando una sequenza CP standard (*cp.av*), con parametri ottimizzati in conformità alle linee guida Bruker per la condizione di Hartmann-Hahn, al fine di massimizzare il trasferimento di magnetizzazione tra ^{1}H e ^{13}C. L’acquisizione è stata eseguita con un numero di scansioni pari a 4000 e 1496 punti acquisiti (TD). Il tempo di acquisizione (AQ) è stato di 32.9 ms, con una larghezza spettrale impostata a 300 ppm. Il tempo di riciclo (RD) è stato impostato a 2.0 s, per garantire un adeguato rilassamento tra le scansioni successive.

3.2 Chemiometria: Workflow convenzionale e AI-driven

In questo paragrafo vengono presentate le strategie adottate per l'elaborazione dei dati metabolomici ottenuti tramite spettroscopia NMR. A tal fine, sono stati sviluppati e applicati due approcci paralleli. Il primo, definito *workflow convenzionale*, si basa sull'utilizzo di strumenti consolidati per il processamento e l'analisi statistica dei dati NMR, nonché per l'individuazione di variabili discriminanti tra classi di campioni, secondo una procedura tradizionalmente impiegata in ambito chemiometrico. Il secondo approccio, invece, è un *workflow AI-driven*, interamente sviluppato in linguaggi di programmazione R e Python e utilizzato per processamento dei dati, classificazione predittiva e identificazione di variabili discriminanti. Questo flusso di lavoro ha sfruttato alcune delle capacità analitiche offerte dall'intelligenza artificiale ed è stato concepito per essere robusto, riproducibile e scalabile.

3.2.1 Workflow classico: TopSpin, AMIX, XLSTAT

3.2.1.1 Elaborazione spettrale NMR con TopSpin

I dati NMR acquisiti sotto forma di Free Induction Decay (FID), ovvero segnali nel dominio del tempo che descrivono il decadimento della magnetizzazione trasversale dei nuclei risonanti, sono stati sottoposti ad apodizzazione esponenziale (Line Broadening di 0.6 Hz) al fine di ridurre il rumore di fondo e migliorare la qualità degli spettri. È stato inoltre impostato il parametro SI (*Size of the Transformed Data*) pari a TD/2 in considerazione dell'acquisizione della sola parte reale del segnale. Successivamente, l'applicazione della trasformata di Fourier ha consentito di convertire i dati nel dominio della frequenza, generando spettri interpretabili, in cui le frequenze di risonanza riflettono la natura e l'ambiente chimico delle specie atomiche presenti. Gli spettri così ottenuti sono stati elaborati manualmente utilizzando il software TopSpin 4.0.2 (Bruker). Le operazioni di processamento hanno previsto la correzione manuale della fase, mediante ottimizzazione dei parametri phC0 e phC1 per l'eliminazione di eventuali distorsioni dello spettro, la correzione della linea di base attraverso l'adattamento a una curva polinomiale e la calibrazione chimica dello spettro utilizzando il riferimento interno TMSPA fissato a 0.00 ppm.

3.2.1.2 Bucketing con AMIX e generazione della matrice dati

Completata l'elaborazione degli spettri, si è proceduto alla segmentazione (*bucketing*) tramite il software AMIX 3.9.15 (Bruker). L'intervallo chimico di interesse, compreso tra 9.60 e −0.08 ppm (escludendo la regione corrispondente al solvente, 4.98–4.72 ppm), è stato suddiviso in

intervalli regolari di 0.04 ppm. Per la scelta dell'ampiezza dell'intervallo di integrazione è stato adottato un valore intermedio, utile per mantenere una buona risoluzione spettrale pur semplificando la struttura dei dati e riducendo la dimensionalità per le analisi statistiche successive (Emwas et al., 2018). Per ciascun intervallo (bucket) è stata calcolata l'area sottesa allo spettro, utilizzando la modalità *No Scaling*, così da preservare integralmente l'informazione originale relativa all'intensità dei segnali. Il risultato di questa procedura è una matrice di dati, nella quale ogni riga rappresenta un campione analizzato e ogni colonna corrisponde a un bucket, ovvero a un intervallo chimico specifico. La matrice finale, composta da 237 variabili (buckets) è stata esportata in formato Excel ed è servita come base per le successive analisi statistiche multivariate ed inferenziali.

3.2.1.3 Analisi con XLSTAT per confronto tra i gruppi e selezione delle variabili significative

Per l'analisi statistica dei dati pre-processati è stato utilizzato il software XLSTAT®, un'estensione avanzata per Excel che consente di eseguire analisi convenzionali, quali PCA e ANOVA, attraverso un'interfaccia accessibile e intuitiva. Questo ha permesso di condurre analisi esplorative e confronti tra le diverse matrici. Inoltre, a partire dai risultati ottenuti, è stata effettuata una selezione mirata delle variabili maggiormente responsabili della separazione osservata tra i gruppi sperimentali. Le sezioni seguenti illustrano le diverse fasi dell'analisi svolta.

Analisi PCA

L'analisi PCA (Principal Component Analysis) in XLSTAT è stata condotta utilizzando le impostazioni di default, basate sulla matrice di correlazione calcolata mediante il coefficiente di Pearson e sulla normalizzazione a n−1. L'esame dello *scree plot* ha consentito di individuare un numero di componenti sufficiente a descrivere adeguatamente la variabilità presente nei dati, attraverso la valutazione del punto di flesso ("elbow method"), corrispondente alla soglia oltre la quale l'incremento di varianza spiegata da componenti aggiuntive risultava marginale.

Le componenti selezionate sono state quindi utilizzate per la costruzione dello *score plot*, che rappresenta i campioni nello spazio delle componenti principali. L'analisi dello score plot ha consentito di valutare eventuali raggruppamenti tra i campioni, fornendo indicazioni preliminari sulla presenza di differenze metaboliche tra le classi analizzate.

Come ultima fase dell'analisi esplorativa, è stata condotta la valutazione dei contributi percentuali alle prime componenti principali (frazioni percentuali dei loadings elevati al

quadrato), al fine di individuare le variabili maggiormente responsabili della separazione osservata tra i gruppi.

ANOVA e test post-hoc di Tukey HSD

Per verificare la significatività statistica delle differenze osservate, è stata condotta un'analisi della varianza (ANOVA) a una via per ciascuna variabile, seguita dall'applicazione del test post-hoc di Tukey HSD per la valutazione dei confronti multipli tra i gruppi.

Sono state selezionate come variabili discriminanti quelle che hanno mostrato un risultato significativo ($p < 0,05$) sia all'ANOVA sia nei confronti multipli evidenziati dal test di Tukey HSD, tra quelle precedentemente individuate sulla base dei contributi percentuali alle componenti principali.

Con riferimento all'analisi esplorativa condotta mediante PCA, tali variabili sono state successivamente prese in esame per l'individuazione di possibili biomarcatori responsabili della separazione osservata tra i gruppi.

Assegnazione delle variabili discriminanti ai metaboliti

Sfruttando i dati disponibili nella letteratura scientifica, nonché le informazioni presenti nei database metabolomici online, tra cui HMDB e BMRB, sono state predisposte specifiche librerie in formato Excel per ciascuna matrice analizzata, ognuna delle quali contenente un elenco di assegnazioni spettrali organizzate per valore di chemical shift. Successivamente, mediante un'operazione di matching tra i bucket corrispondenti alle variabili risultate significative e le librerie di riferimento, è stato possibile associare ciascuna variabile a uno o più metaboliti candidati.

3.2.2 Workflow AI-driven: NMR Workflow, R, Python

Parallelamente al workflow classico, è stato sviluppato un approccio automatizzato, articolato in due passaggi principali:

(1) processamento degli spettri e generazione della matrice di dati (bucketing) mediante script in R;

(2) sviluppo di script in Python per l'analisi supervisionata dei dati mediante modelli di machine learning (dopo specifici passaggi di addestramento e validazione) con successiva individuazione delle variabili discriminanti per la classificazione, finalizzata all'assegnazione dei segnali spettrali ai relativi metaboliti.

Questa architettura ha permesso di ridurre l'intervento manuale, migliorare l'efficienza operativa e garantire una maggiore robustezza dei risultati. Al fine di valutare l'affidabilità complessiva dell'approccio automatizzato, i risultati ottenuti in termini di selezione delle variabili discriminanti sono stati confrontati con quelli derivanti dal workflow convenzionale. Nel seguito vengono descritte in dettaglio le principali fasi del workflow.

3.2.2.1 Processamento degli spettri con NMR Workflow e script in R

Per effettuare il processamento degli spettri in maniera automatizzata, è stato sviluppato uno script in ambiente R in collaborazione con l'azienda Farzati S.p.A., impiegando il pacchetto Rnmr1D, integrato con funzioni personalizzate e supporto al calcolo parallelo, al fine di ottimizzare i tempi di elaborazione. Lo script si è basato sulle fasi procedurali implementate nel tool open-source NMR Workflow, capace di svolgere le principali operazioni di processamento e di fornire un log dettagliato dei diversi passaggi, offrendo così una base concettuale utile per comprendere la logica di sviluppo del flusso di lavoro automatizzato.

Attraverso lo script in R sono state eseguite:

- l'importazione dei dati spettroscopici raw (file 1r);
- il processamento automatizzato dei dati importati, comprendente la calibrazione della scala chimica (ppm) e la correzione globale della linea di base;
- la segmentazione spettrale mediante bucketing uniforme.

Le operazioni sono state eseguite mediante funzioni personalizzate, progettate per adattarsi alle caratteristiche specifiche dei segnali NMR, tenendo conto della dimensione spettrale e dei livelli di rumore stimati.

In particolare, per la calibrazione della scala chimica (asse del chemical shift, δ) è stata sviluppata una funzione che rileva automaticamente il picco del TMSPA (trimetilsililpropionato di sodio), standard interno convenzionalmente fissato a 0.00 ppm. La funzione esegue la ricerca

del massimo locale di intensità all'interno di una finestra centrata sul valore atteso del segnale di riferimento, compresa tra –0.5 e +0.5 ppm. Lo scostamento tra la posizione osservata del picco di riferimento e il valore teorico viene calcolato e applicato come fattore di traslazione lungo l'asse chimico (δ), riallineando correttamente l'intero spettro. L'entità dello spostamento da applicare è modulata tenendo conto anche della risoluzione spettrale, derivata dal numero di punti elaborati (SI), e del livello di rumore presente nello spettro, stimato attraverso la funzione *C_estime_sd*, inclusa nel pacchetto *Rnmr1D*.

La correzione della linea di base (*baseline correction*) è stata effettuata attraverso un algoritmo iterativo di stima locale, già implementato nel pacchetto Rnmr1D. È stata utilizzata in particolare la funzione *C_Estime_LB()*, che applica una finestra mobile lungo lo spettro e stima l'altezza della linea di base in ciascuna regione, adattando dinamicamente il livello di correzione alla forma locale del segnale. Il grado di intervento dell'algoritmo è stato regolato tramite il parametro GBCLEV, che consente di selezionare tra diversi livelli di intensità della correzione. Per il presente lavoro, è stato adottato un livello soft (GBCLEV = 1), sufficiente a correggere le lievi deviazioni osservate negli spettri senza rischiare alterazioni dei segnali reali a bassa intensità.

Per l'implementazione della fase di segmentazione spettrale (bucketing) è stato sviluppato uno script dedicato, in grado di eseguire automaticamente tutte le operazioni necessarie alla discretizzazione dello spettro NMR. In particolare, lo script identifica la regione spettrale di interesse (compresa tra 9.60 e −0.08 ppm, con esclusione dell'intervallo 4.90–4.70 ppm corrispondente al segnale residuo del solvente), la suddivide in intervalli uniformi di ampiezza pari a 0.04 ppm e calcola, per ciascun intervallo, il rapporto segnale/rumore (SNR). Questo passaggio consente di selezionare esclusivamente i bucket contenenti segnali metabolici significativi, escludendo quelli privi di informazione utile. Infine, per ciascun bucket selezionato, vengono determinate le intensità spettrali e normalizzate secondo il metodo CSN (Constant Sum Normalization), ottenendo una matrice di dati adatta alle successive analisi.

L'intero blocco di preprocessing ha richiesto tempi di ottimizzazione più lunghi del previsto, a causa della complessità delle operazioni coinvolte e della necessità di costruire diverse funzioni su misura. L'obiettivo era realizzare una pipeline flessibile, adattabile a matrici diverse e in grado di garantire coerenza e accuratezza.

Per questo motivo, e nell'ottica di ottimizzare il flusso di lavoro complessivo, si è scelto di adottare la matrice di buckets generata tramite il software AMIX, già validata e consolidata nell'ambito del protocollo convenzionale.

3.2.2.1 Sviluppo di script in Python per l'analisi predittiva e l'identificazione delle variabili discriminanti

Il workflow automatizzato (AI-driven) per l'analisi dei dati spettrali è stato progettato con l'obiettivo di garantire rigore metodologico, riproducibilità dei risultati e riduzione dell'intervento manuale. L'approccio è stato implementato mediante script in linguaggio Python su piattaforma Google Colab, strutturato in moduli sequenziali che coprono l'intero processo analitico: dalla gestione iniziale dei dati fino all'attribuzione spettrale dei segnali selezionati. Nel seguito vengono descritte le principali fasi operative del workflow.

Il punto di partenza è rappresentato da una matrice di dati in formato .csv, strutturata con una prima colonna identificativa dei campioni (replica), seguita da 230 variabili numeriche (Var1–Var230) corrispondenti a intervalli spettrali da 0 a 10 ppm, con passo regolare di 0.04 ppm, ottenuti dopo l'eliminazione del segnale dello standard interno. L'ultima colonna (Class) riporta l'etichetta di appartenenza di ciascun campione al gruppo sperimentale di riferimento, variabile in funzione della matrice analizzata (ad esempio "Norberto" per la varietà di frumento, "Arborea_IT" per la tipologia di latte o "Italia" per l'origine geografica del pomodoro).

Fasi operative del workflow

Importazione, verifica e pre-processing dei dati

I dati contenuti nella matrice dei buckets sono stati importati e sottoposti a una verifica preliminare di integrità e completezza, al fine di assicurare la qualità del dataset per le successive fasi analitiche. Successivamente, è stata applicata una procedura di scalatura mediante standardizzazione Z-score, finalizzata a uniformare la varianza delle variabili (con ottenimento di distribuzioni a media nulla e deviazione standard unitaria) e a ottimizzare le prestazioni degli algoritmi di apprendimento automatico. All'interno dello script sono stati inoltre implementati test di normalità (Kolmogorov-Smirnov e Shapiro-Wilk) per valutare la distribuzione delle variabili standardizzate, al fine di verificare il rispetto delle assunzioni statistiche richieste per le successive analisi multivariate.

Un estratto dello script utilizzato per la lettura dei dati contenuti nella matrice dei buckets e la successiva standardizzazione è riportato in Appendice.

Riduzione dimensionale mediante PCA

Al fine di affrontare efficacemente l'elevata dimensionalità del dataset e contenere il rischio di overfitting, si è proceduto alla trasformazione delle variabili originarie in un nuovo insieme di componenti latenti (PCs), ottenute tramite Principal Component Analysis (PCA) (Corsaro et al., 2022; Yata & Aoshima, 2012; Liakos et al., 2018). Questa trasformazione ha consentito di proiettare i dati in uno spazio a dimensionalità ridotta, preservando la massima quantità possibile di varianza informativa. La selezione del numero ottimale di componenti principali da mantenere è stata effettuata integrando tre criteri complementari: la soglia del 95% di varianza cumulativa spiegata, il criterio di Kaiser (conservazione delle componenti con autovalori superiori a 1) e il metodo del Broken Stick, basato sul confronto con una distribuzione casuale teorica (Jolliffe & Cadima, 2016; Peres-Neto et al., 2005; Stephens, 1974). L'applicazione congiunta di questi criteri ha garantito una selezione robusta delle componenti latenti, riducendo la dimensionalità del dataset senza comprometterne il contenuto informativo essenziale.

Un estratto dello script dedicato all'esecuzione della PCA è riportato in Appendice.

Ottimizzazione e addestramento dei modelli di machine learning supervisionato

A valle della PCA, le coordinate dei campioni proiettate nello spazio delle componenti latenti sono state utilizzate per testare il funzionamento di diversi modelli di classificazione supervisionata, selezionati sulla base di una revisione della letteratura scientifica che ne attesta l'efficacia in presenza di dataset caratterizzati da elevata complessità e ridotta numerosità campionaria (Bifarin, 2023; Galal et al., 2022; Pomyen et al., 2020). I modelli implementati comprendono: Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes, Decision Tree, Random Forest e XGBoost.

Per incrementare la robustezza dei classificatori e simulare realistiche fluttuazioni sperimentali, ai dati di input è stato aggiunto un rumore gaussiano controllato, con media zero e deviazione standard σ compresa tra 0,05 e 0,1 (Ljunggren, n.d.).

In una prima fase, si è proceduto all'ottimizzazione dei modelli, finalizzata alla selezione dei valori ottimali di una serie di iperparametri che ne regolano il comportamento interno e incidono direttamente sulle prestazioni predittive.

Nella tabella seguente sono riportati i principali iperparametri considerati per ciascun modello.

Tabella 3.3 – Iperparametri degli algoritmi di classificazione

Modello	**Iperparametro**	**Significato**	**Valori testati**
Logistic Regression	C	Inverso della forza di regolarizzazione	[0.001, 0.01, 0.1, 1, 10]
	penalty	Tipo di regolarizzazione	['l1', 'l2', 'elasticnet']
	solver	Algoritmo di ottimizzazione	['liblinear', 'saga', 'lbfgs']
Naïve Bayes	var_smoothing	Fattore di stabilizzazione per evitare divisioni per zero	[da 1e-9 a 1e-6]
k-NN	n_neighbors	Numero di vicini considerati	[1–20]
	weights	Modalità di ponderazione dei vicini	['uniform', 'distance']
	metric	Tipo di distanza utilizzata	['euclidean', 'manhattan']
SVM	C	Penalizzazione degli errori di classificazione	[0.01, 0.1, 1, 10]
	kernel	Funzione kernel utilizzata	['linear', 'rbf', 'poly']
	gamma	Coefficiente del kernel (solo per 'rbf' e 'poly')	[0.001, 0.01, 0.1, 1]
Decision Tree	max_depth	Profondità massima dell'albero	[None, 3, 5, 10, 20]
	min_samples_split	Numero minimo di campioni per effettuare uno split	[2, 5, 10]
	min_samples_leaf	Numero minimo di campioni in una foglia	[1, 2, 5]
	max_features	Numero massimo di feature da considerare per ogni split	['auto', 'sqrt', 'log2']
	criterion	Funzione per misurare la qualità dello split	['gini', 'entropy']
Random Forest	n_estimators	Numero di alberi nella foresta	[100–500]
	max_depth	Profondità massima degli alberi	[5–30]
	min_samples_split	Numero minimo di campioni per effettuare uno split	[2–20]
	max_features	Numero massimo di feature per split	['sqrt', 'log2', None]
XGBoost	n_estimators	Numero di alberi	[100–1000]
	learning_rate	Tasso di apprendimento	[0.01–0.3]
	max_depth	Profondità massima degli alberi	[3–10]
	subsample	Frazione di campioni utilizzati per ogni albero	[0.5–1.0]
	colsample_bytree	Frazione di feature utilizzate per ogni albero	[0.5–1.0]
	gamma, lambda, alpha	Parametri di regolarizzazione per controllare la complessità del modello	[0, 0.1, 0.5, 1]

Il valore ottimale degli iperparametri per ciascun algoritmo è stato determinato tramite la funzione GridSearchCV della libreria *scikit-learn*, che consente di esplorare in modo sistematico tutte le combinazioni possibili definite all'interno di una griglia predefinita e di identificare automaticamente la configurazione che garantisce le migliori performance predittive. Questo processo si basa su una procedura di validazione incrociata stratificata, integrata all'interno della funzione stessa.

Nel dettaglio, la validazione incrociata prevede la suddivisione del dataset in k sottoinsiemi (fold), con k selezionato in funzione del numero di campioni disponibili (nel presente lavoro sono stati utilizzati valori compresi tra 2 e 5). In ogni iterazione, il modello viene addestrato su $k - 1$ fold e testato sul fold rimanente, ripetendo il processo k volte in modo che ciascun sottoinsieme venga utilizzato una volta come set di test. La stratificazione applicata in ciascuna suddivisione garantisce che la distribuzione delle classi venga mantenuta proporzionale in ogni fold rispetto all'intero dataset, contribuendo a migliorare l'affidabilità della stima delle performance e a ridurre il rischio di distorsioni dovute a sbilanciamenti.

Per ciascuna combinazione di iperparametri, GridSearchCV calcola la media delle accuratezze ottenute nei k test e seleziona come ottimale la configurazione che massimizza tale valore.

Una volta individuati i parametri ottimali, ciascun modello è stato addestrato sull'intero dataset disponibile (comprensivo del rumore), utilizzando la funzione .fit(X, y), che consente al classificatore di apprendere le relazioni tra le variabili predittive (componenti latenti) e le corrispondenti classi di appartenenza.

Al termine dell'addestramento, i modelli sono stati sottoposti a una nuova validazione incrociata stratificata, utilizzando la funzione cross_val_score con lo stesso numero di fold (k) della fase di ottimizzazione. In questo contesto, sono stati calcolati la media e la deviazione standard dell'accuratezza, utilizzando quest'ultima come indicatore della stabilità predittiva del modello.

In parallelo, è stata condotta un'analisi dell'importanza delle variabili predittive mediante la tecnica della *permutation importance*, al fine di quantificare il contributo informativo di ciascuna coordinata (componente latente) alla capacità discriminante del modello.

Al termine del processo, tutti i modelli ottimizzati sono stati salvati in formato .pkl, per garantirne la riutilizzabilità nelle fasi successive del workflow.

Un estratto dello script utilizzato per l'ottimizzazione degli iperparametri, l'addestramento e il salvataggio dei modelli di classificazione, nonché per l'individuazione delle componenti discriminanti, è riportato in Appendice.

Valutazione comparativa delle performance dei modelli ottimizzati

La fase successiva del workflow ha previsto una validazione indipendente e standardizzata dei modelli di classificazione precedentemente ottimizzati e salvati. A tal fine è stato sviluppato uno script dedicato, il quale ha caricato i classificatori memorizzati in formato .pkl - addestrati in presenza di rumore gaussiano controllato - e li ha sottoposti a una procedura di validazione incrociata stratificata a 2 fold, implementata mediante la funzione cross_val_predict. Questa tecnica consente di generare una previsione per ciascun campione del dataset, assicurando che ogni predizione provenga da un modello che non ha avuto accesso a quel campione durante la fase di addestramento. In questo modo, è possibile ottenere una stima oggettiva e realistica della capacità di generalizzazione del modello, riducendo il rischio di overfitting e garantendo condizioni di test indipendenti.

Per ciascun classificatore sono state calcolate metriche di valutazione standard, tra cui l'accuracy, l'F1-score (macro e pesato), la precisione e il richiamo, in accordo con quanto raccomandato in letteratura per l'analisi di classificatori multiclasse (Vujović, 2021).

Questa fase di validazione non ha comportato alcuna modifica o riaddestramento dei modelli, ma ne ha valutato esclusivamente la robustezza predittiva in condizioni sperimentali controllate e replicabili. L'adozione di una pipeline omogenea — per quanto riguarda la configurazione dei dati, il livello di rumore e le metriche di valutazione — ha permesso un confronto diretto e imparziale tra gli algoritmi considerati, facilitando l'identificazione del classificatore più stabile e performante per la matrice in esame.

Un estratto dello script utilizzato per la valutazione comparativa delle prestazioni predittive su dati non precedentemente utilizzati in fase di addestramento è riportato in Appendice.

Ricostruzione delle Variabili Discriminanti

Basandosi sull'individuazione delle componenti principali più rilevanti per la discriminazione tra classi, ottenuta mediante *permutation importance* (equivalente alla *Mean Decrease in Accuracy* – MDA), si è proceduto all'identificazione delle variabili originarie (buckets spettrali) maggiormente responsabili di tali componenti.

A tal fine, è stata implementata una procedura di rimappatura tra l'importanza attribuita alle componenti principali dai modelli predittivi e i contributi percentuali dei buckets spettrali alle stesse, al fine di trasferire l'informazione discriminante dalle feature latenti alle variabili reali del dominio NMR.

Per ciascun modello predittivo è stata selezionata automaticamente una rosa di 20 variabili a maggiore rilevanza. La selezione finale delle variabili discriminanti è stata poi effettuata mediante un approccio basato sul voto di maggioranza, considerando come significative le variabili individuate da un numero consistente di modelli. Tale strategia ha permesso di rafforzare la robustezza del processo di selezione e di valutare la convergenza tra approcci metodologicamente eterogenei.
Un estratto dello script sviluppato per combinare i contributi percentuali delle variabili (ottenuti dall'analisi PCA) con l'importanza delle componenti principali calcolata tramite permutation importance derivante dai modelli predittivi è riportato in Appendice. Viene anche mostrato il codice per il salvataggio del file finale contenente l'importanza pesata di ciascuna variabile, distinta per ogni modello.

Validazione mediante set di dati esterno

Per simulare un contesto operativo di predizione reale, è stato utilizzato un set di dati privo di rumore artificiale come test esterno. I modelli addestrati sono stati applicati a questo dataset per valutare la capacità predittiva effettiva sui dati non perturbati, misurata tramite il calcolo di metriche convenzionali compresa la matrice di confusione. Al fine di garantire l'uniformità del flusso di elaborazione, i dati del set di test sono stati preprocessati utilizzando il modello di standardizzazione e la trasformazione PCA precedentemente impiegati durante la fase di training. In questo modo si è garantita la coerenza tra le fasi di addestramento e simulazione.
Un estratto dello script sviluppato per il test di predizione dei modelli applicato a dati del set esterno è riportato in Appendice.

L'integrazione dei diversi moduli descritti finora ha consentito di costruire un flusso di lavoro completo, efficiente e facilmente adattabile a differenti matrici spettrali. I risultati ottenuti con l'approccio automatizzato sono stati successivamente confrontati con quelli derivanti dal workflow classico, al fine di valutarne l'affidabilità, l'efficacia nella selezione delle variabili discriminanti e la coerenza nell'interpretazione metabolomica.

3.3 Integrazione tra analisi spettroscopica HR-MAS NMR e Blockchain: proposta metodologica e implementazione operativa

Di seguito viene presentata una proposta metodologica per l'integrazione della tecnologia blockchain all'interno del workflow sperimentale e analitico descritto nel presente lavoro. L'approccio si basa sull'impiego di un'architettura blockchain permissioned (es. Hyperledger Fabric), con l'obiettivo di garantire l'integrità, la tracciabilità e l'immutabilità dei dati spettrali NMR acquisiti da matrici agroalimentari.

L'idea fondante è quella di trasformare la bio-impronta metabolica del campione - rappresentata dallo spettro 1H HR-MAS NMR e dai relativi metadati - in un passaporto digitale verificabile, registrato su un sistema distribuito. Questo passaporto digitale consente di associare in modo univoco una impronta spettrale ad un campione fisico, supportando sistemi di certificazione basati su evidenze scientifiche oggettive, e non più esclusivamente su documentazione cartacea o autodichiarazioni.

3.3.1 Struttura metodologica proposta

Il flusso metodologico si articola in quattro fasi principali:

1. **Acquisizione degli spettri NMR**

   I campioni agroalimentari vengono analizzati mediante spettroscopia 1H HR-MAS NMR. I dati grezzi (FID) sono salvati nel formato nativo Bruker, corredati dai relativi metadati tecnici (strumento, operatore, data, temperatura, sequenza di impulso, ecc.).

2. **Serializzazione e generazione dell'hash**

   Prima di qualsiasi elaborazione analitica, ciascun FID viene convertito (serializzato) in un file strutturato in formato JSON, contenente:

   - identificativo univoco del campione,
   - metadati sperimentali,
   - scala chimica in ppm,
   - array di intensità reali (FID),
   - timestamp di generazione.

   A partire dal contenuto completo del file JSON viene quindi calcolato un hash crittografico (SHA256), che rappresenta l'identificatore digitale univoco e immutabile del campione.

3. **Registrazione su rete blockchain**

   Lo SHA256 generato, insieme all'identificativo del campione, ai metadati e al timestamp, può essere registrato tramite smart contract su una rete blockchain permissioned. Questo consente di garantire:

   - tracciabilità del dato,
   - integrità rispetto a modifiche non autorizzate,
   - verifica futura del contenuto originario.

4. **Elaborazione e verifica**

   Una volta registrati su blockchain, i dati possono essere elaborati attraverso i workflow automatizzati descritti nei paragrafi precedenti (PCA, modelli predittivi AI-driven). In qualunque fase successiva, è possibile confrontare un file spettrale con il suo hash originario per verificarne l'autenticità.

3.3.2 Implementazione operativa (Google Colab)

Per dimostrare la fattibilità tecnica di questa integrazione, è stato realizzato uno script Python eseguibile in Google Colab (riportato in Appendice), che consente di:

- leggere un file FID Bruker (fid) a partire dalla directory contenente i file di acquisizione,
- estrarre e strutturare metadati e informazioni sperimentali,
- generare un file JSON contenente intensità, ppm scale e metadati,
- calcolare l'hash SHA256 del file JSON.

3.4 Integrazione tra analisi spettroscopica CP-MAS NMR e sistemi di tracciabilità basati su sensori IoT e Blockchain

Nell'ambito della collaborazione con l'azienda Farzati S.p.A., è stata concettualmente definita una piattaforma comune per l'integrazione delle informazioni provenienti da analisi spettroscopiche NMR e da sensoristica IoT, con registrazione dei dati su rete blockchain. L'attività ha riguardato l'analisi di campioni di legno cippato, appartenenti a diverse specie arboree e raccolti in differenti aree geografiche. I tecnici incaricati hanno utilizzato il dispositivo portatile BluDev®, sviluppato da Farzati S.p.A., per registrare, al momento del prelievo, la posizione geografica dei campioni tramite il modulo GPS integrato, acquisendo contestualmente una serie di parametri fisico-chimici mediante spettroscopia del vicino infrarosso (NIR), anch'essa integrata nel sensore.

In Figura 3.2 è riportato, a titolo esemplificativo, la localizzazione geografica acquisita con il sensore BluDev® per uno dei campioni analizzati (Lat. 44.650352, Lon. 11.891272).

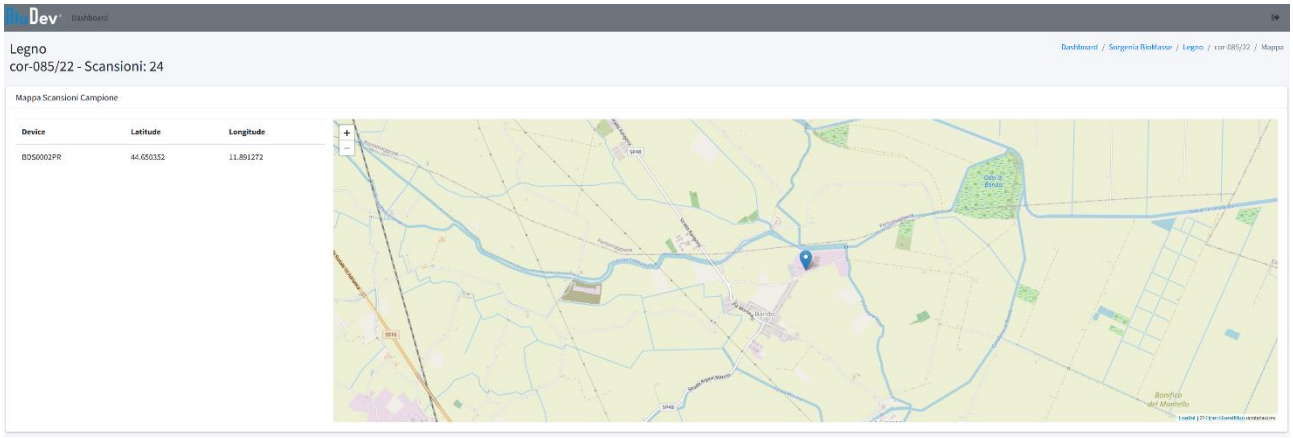


Figura 3.2 - Geolocalizzazione del campione BDS0002PR tramite sensore BluDev®

I dati raccolti con il sensore sono stati automaticamente criptati e registrati su una rete blockchain, garantendo l'immutabilità e la tracciabilità delle informazioni, ed è stato generato un QR code associato a ciascun campione.

Successivamente, i campioni tracciati sono stati trasferiti in laboratorio e sottoposti ad analisi mediante spettroscopia 13C CP-MAS NMR. Le condizioni sperimentali di acquisizione, inclusi i parametri strumentali e le modalità di elaborazione degli spettri, sono dettagliatamente descritte nel Capitolo 3, paragrafo 3.1.2.2. Tali condizioni hanno consentito di acquisire spettri ad alta risoluzione e buona riproducibilità, offrendo una visione dettagliata e affidabile della composizione carboniosa della biomassa (Duchesne et al., 2001; Gao et al., 2015).

In una prima fase è stata effettuata una caratterizzazione generale dei cippati sulla base delle regioni spettrali e dei gruppi funzionali identificati.

Per approfondire ulteriormente la variabilità compositiva, è stata poi eseguita un'analisi gerarchica basata sulle percentuali relative dei diversi gruppi carboniosi, ottenuti dall'analisi 13C CP-MAS NMR, i cui risultati sono stati rappresentati attraverso dendrogrammi.

Infine, per supportare l'interpretazione complessiva delle variazioni nella composizione tra i campioni, è stata condotta un'analisi delle componenti principali (PCA), basata sulla distribuzione percentuale dei gruppi carboniosi identificati.

I dati ottenuti mediante spettroscopia NMR sono stati correlati con le informazioni raccolte in campo tramite il sensore BluDev®, consentendo un'analisi integrata della qualità chimica e dell'origine geografica dei materiali.

Capitolo 4 Risultati e discussione

Il presente capitolo è dedicato alla presentazione e discussione dei risultati ottenuti attraverso gli approcci descritti nel Capitolo 3. Nei paragrafi dal 4.1 al 4.4, sono riportati i risultati relativi all'analisi delle matrici agroalimentari considerate, ovvero *Triticum durum – Triticum monococcum*, *Triticum durum* varietà "Senatore Cappelli", concentrato di pomodoro e latte UHT. Inizialmente viene presentata la descrizione dei profili spettrali ottenuti mediante spettroscopia 1H HR-MAS NMR per le diverse classi di campioni. Successivamente, rispettando la suddivisione metodologica adottata, sono presentati in sequenza i risultati relativi al metodo convenzionale, basato sull'applicazione di tecniche statistiche consolidate per l'analisi degli spettri NMR, e al metodo AI-driven, che ha previsto l'integrazione di algoritmi di machine learning e strategie automatizzate di elaborazione dei dati. Con riferimento all'approccio convenzionale, sono dapprima illustrati i risultati relativi all'identificazione dei raggruppamenti mediante analisi delle componenti principali, seguiti dalla selezione delle variabili maggiormente significative per la discriminazione tra i gruppi e, infine, dall'attribuzione ai metaboliti potenzialmente rilevanti. In relazione all'approccio AI-driven, sono dapprima riportati i risultati relativi alla selezione delle componenti latenti, seguiti dalla sintesi delle performance predittive dei modelli sviluppati e, infine, dall'individuazione delle variabili discriminanti tra le classi, con la corrispondente identificazione dei metaboliti associati. Vengono inoltre riportati i dati relativi alla performance dei modelli di machine learning nella simulazione del processo predittivo, effettuata utilizzando un set di dati esterno costituito dai dati originali privi di rumore. Al termine dei paragrafi, viene proposta una discussione finalizzata a dimostrare la coerenza tra i due approcci (convenzionale e AI-driven), attraverso la valutazione della convergenza delle variabili discriminanti selezionate e dell'affidabilità dei metaboliti identificati, quest'ultima considerata in funzione della loro capacità di supportare l'interpretazione biologica delle differenziazioni tra le classi.

Il paragrafo 4.5 è dedicato ai risultati ottenuti dall'analisi dei campioni di cippato. A partire da materiali georeferenziati tramite sensore BluDev® e registrati in blockchain, vengono presentati i profili di composizione lignocellulosica ottenuti mediante spettroscopia 13C CP-MAS NMR allo stato solido, accompagnati dai risultati delle analisi statistiche multivariate finalizzate a esplorare le relazioni tra composizione chimica, origine botanica e provenienza geografica dei campioni.

4.1 Triticum durum e triticum monococcum

4.1.1 Descrizione dei profili spettrali

Gli spettri 1H-NMR ottenuti per le diverse classi di *Triticum durum* e *Triticum monococcum* sono riportati in Figura 4.1. Nella regione spettrale compresa tra δ 0.8 e 2.5 ppm si osservano segnali riconducibili ad amminoacidi a catena ramificata, acidi grassi e composti alifatici (Romano et al., 2022). Tali segnali, generalmente ben definiti, risultano tuttavia soggetti a variazioni di intensità tra le classi, riflettendo potenziali differenze nella composizione proteica e nel grado di maturazione della cariosside (Shewry et al., 2017). La regione δ 3.0–5.5 ppm, solitamente la più ricca di segnali, è dominata dalla presenza di zuccheri solubili (glucosio, fruttosio, saccarosio), fruttani e alcoli zuccherini. In quest'area non si evidenziano differenze marcate tra le classi. Infine, nella regione aromatica compresa tra δ 6.0 e 9.5 ppm – meno intensa in termini di segnali, ma metabolicamente informativa – sono stati rilevati picchi riconducibili a composti fenolici, tra cui acidi idrossicinnamici e derivati polifenolici (Shewry et al., 2017b). Queste molecole, spesso coinvolte nei meccanismi di difesa delle piante, possono fornire informazioni rilevanti sulla varietà, sull'ambiente di coltivazione e sulla risposta agli stress abiotici.

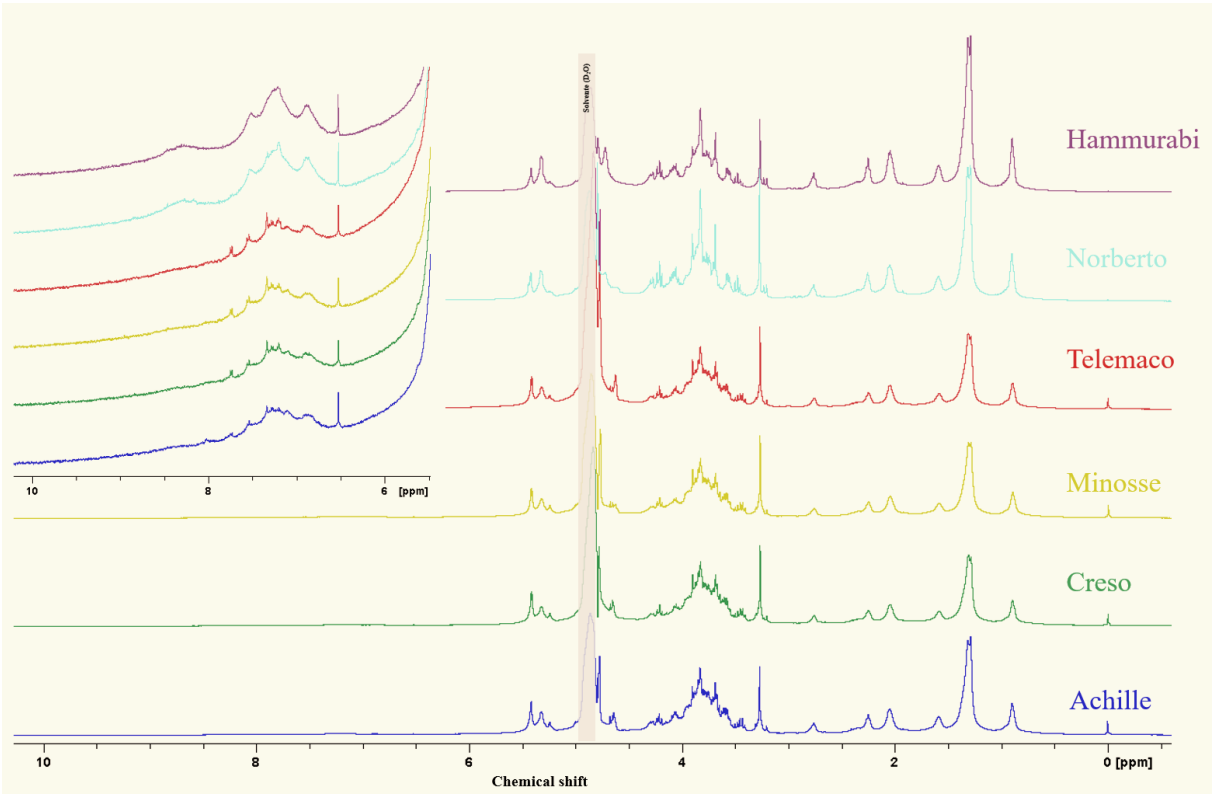


Figura 4.1 - Triticum durum e triticum monococcum - Spettri 1H-NMR.

4.1.2 Risultati dell'analisi chemiometrica convenzionale

PCA

La Figura 4.2 mostra due esempi significativi di score plot ottenuti dall'analisi PCA in XLSTAT. Nello score plot PC1-PC2 vi è una chiara separazione delle varietà Hammurabi e Norberto rispetto alle altre, evidenziata lungo la prima componente principale (PC1). Questa separazione è verosimilmente riconducibile alla differente appartenenza tassonomica: le varietà Hammurabi e Norberto appartengono difatti alla specie Triticum monococcum (farro monococco), mentre le restanti varietà analizzate sono riconducibili a Triticum durum (frumento duro).

L'analisi nel piano definito dalle componenti principali PC1 e PC4 rivela ulteriori elementi distintivi nella distribuzione dei campioni. Questa proiezione consente di esplorare con maggiore dettaglio le relazioni tra le varietà di frumento duro. La varietà Achille mostra una distribuzione relativamente compatta lungo PC4, suggerendo una certa coerenza metabolica interna. Al contrario, Telemaco presenta una dispersione più ampia nello spazio delle componenti, indicativa di una maggiore eterogeneità tra i campioni. Creso tende a occupare una regione prossima a quella di Telemaco, mentre Minosse si colloca in una posizione intermedia, mostrando parziali sovrapposizioni con entrambe. Nel complesso, la configurazione osservata suggerisce l'esistenza di un gradiente di similarità metabolica tra le varietà moderne di frumento duro, associato a diversi livelli di omogeneità interna.

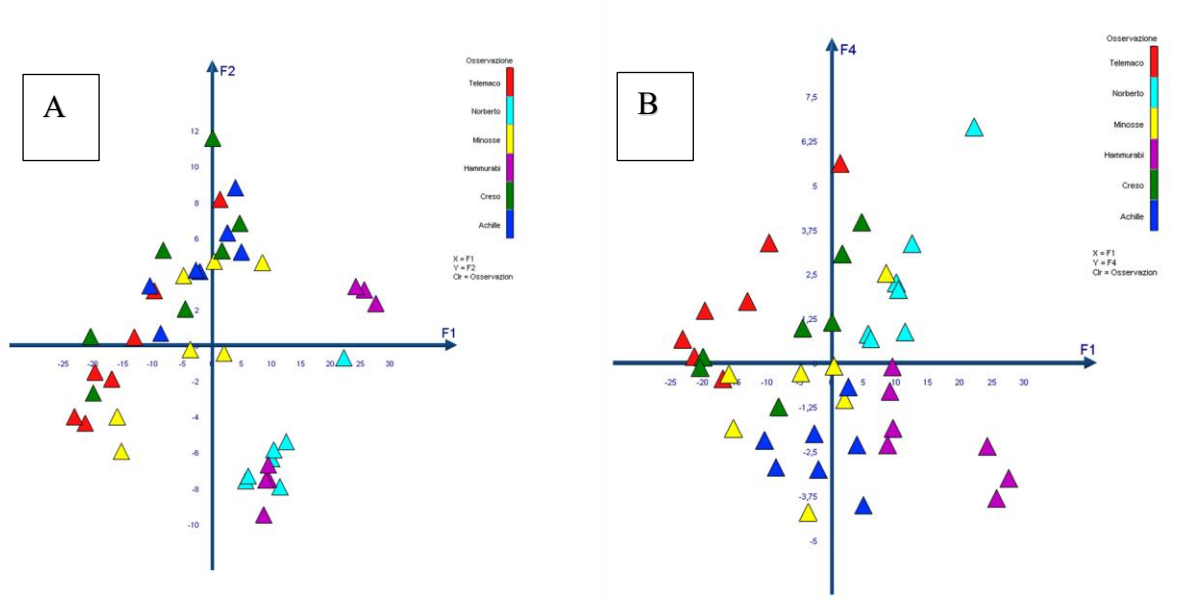


Figura 4.2 – Triticum durum e triticum monococcum – Score Plot da PCA.

Norberto e Hammurabi, appartenenti entrambi al farro monococco (Triticum monococcum), risultano parzialmente separati nello spazio definito da PC1 e PC4. Tale evidenza indica che, pur condividendo la medesima appartenenza tassonomica, queste due varietà esprimono profili metabolici distinti, potenzialmente riconducibili a differenze genetiche varietali o a pratiche agronomiche differenziate (Augustijn et al., 2021).

Selezione delle variabili significative

La Tabella 4.1 riporta, per ciascuna delle prime cinque componenti principali, l'elenco delle variabili che risultano significativamente differenti tra almeno due gruppi ($p < 0.05$), sulla base dell'analisi della varianza univariata (ANOVA). È opportuno precisare che la significatività statistica indicata non implica che ciascuna variabile discrimini tutte le classi tra loro, bensì che, per almeno un confronto tra coppie di gruppi, è stata rilevata una differenza significativa.
Per una maggiore chiarezza interpretativa, è stata successivamente condotta un'analisi post-hoc (test di Tukey HSD) al fine di individuare con precisione quali gruppi differiscono tra loro per ciascuna variabile significativa. Nella presente tabella è tuttavia riportato esclusivamente l'elenco complessivo delle variabili che hanno mostrato una significatività globale (ANOVA a $p < 0.05$), senza entrare nel dettaglio dei confronti specifici tra classi. Tale approccio è stato adottato con l'obiettivo di fornire una sintesi preliminare delle variabili maggiormente coinvolte nella separazione dei gruppi lungo ciascuna componente principale.

Tabella 4.1 - Triticum durum e triticum monococcum - Elenco delle variabili significative.

	Selezione PCA+ANOVA+Tukey
PC	Variabili significative ($p < 0.05$)
PC1	Var36, Var37, Var49, Var50, Var51, Var52, Var53, Var54, Var55, Var56, Var62, Var63, Var64, Var65, Var66, Var67, Var68, Var69, Var70, Var71
PC2	Var103, Var105, Var109, Var136, Var144, Var145, Var150, Var199, Var200, Var201
PC3	Var1, Var135, Var136, Var137, Var138, Var139, Var140, Var141, Var142, Var143, Var144, Var145, Var146, Var147, Var148, Var149, Var150, Var151, Var152, Var153
PC4	Var152, Var153, Var154, Var204, Var205, Var214, Var215, Var216, Var217, Var218, Var219, Var220, Var221, Var222, Var223, Var224, Var225, Var226, Var227, Var228, Var229, Var230, Var138, Var139
PC5	Var140, Var141, Var142, Var143, Var146, Var156, Var157, Var158, Var159, Var224, Var225, Var226, Var227, Var228, Var229, Var230

Assegnazione delle variabili discriminanti ai metaboliti

La Tabella 4.2 seguente riporta un set selezionato di bucket spettrali (1H NMR), risultati statisticamente significativi nel confronto tra varietà di frumento antico e moderno. Le variabili sono state selezionate sulla base della loro rilevanza nei modelli PCA e successivamente assegnate, in modo putativo, a metaboliti noti attraverso l'osservazione dei segnali 2D HSQC e il confronto con librerie spettrali.

Tabella 4.2 - Triticum durum e triticum monococcum - Assegnazione putativa delle variabili ai metaboliti.

Buckets	PC	δ ^{1}H (ppm)	δ ^{13}C (ppm)	Metabolita Assegnato
Var105	PC2	5,41	102,99	Saccarosio
Var103	PC2	5,24	94,81	Glucosio
Var144	PC2, PC3	3,83	63,19	Serina
Var145	PC2, PC3	4,21	63,39	Fruttosio
Var148	PC3	3,71	75,25	Glucosio
Var146	PC3, PC5	3,64	79,67	Glucosio (
Var149	PC3	3,59	73,84	Glucosio
Var150	PC2, PC3	3,45	71,83	Glucosio)
Var152	PC3, PC4	3,27	56,16	Etanolamina / Colina
Var199	PC2	1,59	27,65	Alanina

Le componenti PC2 e PC3 evidenziano in particolare segnali legati al metabolismo degli zuccheri semplici, come glucosio, fruttosio e saccarosio, riflettendo potenziali differenze nel profilo carboidratico tra i gruppi varietali. La presenza simultanea di più segnali riconducibili al glucosio (Var103, Var146, Var148, Var149, Var150) suggerisce che il contenuto e la forma di questo zucchero possano contribuire in modo rilevante alla separazione delle varietà.

All'interno della PC3, compaiono inoltre metaboliti a contenuto azotato come la serina (Var144) e l'etanolamina o colina (Var152), potenzialmente associabili a differenze nei profili amminoacidici e nei precursori di composti funzionali. Infine, nella PC2 è stato identificato un doppio segnale per l'alanina (Var199), in due forme spettrali lievemente diverse, che potrebbe indicare un coinvolgimento di questo amminoacido nei meccanismi di risposta o adattamento metabolico.

4.1.3 Risultati dell'analisi chemiometrica AI-driven

Importazione, verifica e pre-processing dei dati

La Figura 4.3 mostra un estratto della matrice dei bucket spettrali importata nello script, rappresentando i dati prima e dopo la standardizzazione mediante trasformazione Z-score.
Tale standardizzazione realizza una distribuzione centrata attorno allo zero e priva di dominanze numeriche, condizione ottimale per l'applicazione di modelli di machine learning supervisionato. Questo passaggio garantisce che ogni variabile contribuisca in egual misura al modello, prevenendo distorsioni dovute alla diversa scala delle intensità.

Dati Grezzi (prime 5 righe):

	replica	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	...	Var222	
0	1	423389.5870	447691.5377	440382.0620	481812.2332	495494.2497	515468.6788	551592.0767	564360.9538	597702.1037	...	4895111.440	4433
1	2	367777.1575	396266.8846	425006.3666	414994.6540	472889.2673	501298.0728	501858.4379	534819.0206	569576.4830	...	5292347.823	4835
2	3	428321.4733	438357.1108	436315.3860	448799.0641	488740.5389	502710.0134	533636.5646	559268.3059	574638.7932	...	4147698.728	3734
3	4	454851.1899	486771.7636	501712.4307	527611.4102	536807.7385	576851.6696	602167.5245	642279.3090	677948.5357	...	5767436.079	5271
4	5	404259.1439	449681.4307	441258.7339	485249.3791	526688.3769	545474.8321	574130.2899	599171.9752	628242.6098	...	5508812.431	4986

5 rows × 232 columns

Dati Standardizzati (prime 5 righe):

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224
0	-0.108612	-0.098833	-0.285651	-0.136938	-0.178410	-0.215906	-0.157432	-0.219247	-0.184526	-0.239067	...	0.205877	0.253452	0.231999	0.236713
1	-0.561416	-0.495351	-0.403608	-0.628740	-0.339500	-0.313757	-0.478972	-0.405847	-0.354269	-0.445547	...	0.675049	0.713206	0.747467	0.744643
2	-0.068456	-0.170808	-0.316849	-0.379927	-0.226539	-0.304007	-0.273519	-0.251415	-0.323717	-0.282601	...	-0.640557	-0.611590	-0.662279	-0.643183
3	0.147553	0.202501	0.184858	0.200160	0.116003	0.207955	0.169551	0.272920	0.299775	0.197855	...	1.219910	1.263064	1.305190	1.303505
4	-0.264375	-0.083490	-0.278925	-0.111639	0.043889	-0.008708	-0.011717	0.000635	-0.000209	0.056977	...	0.943389	0.963738	0.940303	0.987215

5 rows × 230 columns

Figura 4.3 – – Triticum durum e triticum monococcum - Estratto della matrice dei bucket spettrali prima e dopo la standardizzazione Z-score Matrice dei buckets importata prime e dopo la standardizzazione Z-score

Riduzione dimensionale mediante PCA

In analogia con quanto mostrato nel paragrafo 4.1.2, vengono presentati i risultati dell'analisi PCA condotta con l'impiego dello script. La Figura 4.4 riporta l'andamento della varianza cumulativa spiegata (Elbow Plot), evidenziando come la soglia del 95% venga raggiunta entro le prime cinque componenti principali. La combinazione dei criteri adottati — soglia del 95% della varianza spiegata, criterio di Kaiser e metodo del "Broken Stick" — ha condotto alla selezione delle prime cinque componenti principali, ritenute ottimali per l'addestramento e la predizione dei modelli di machine learning.

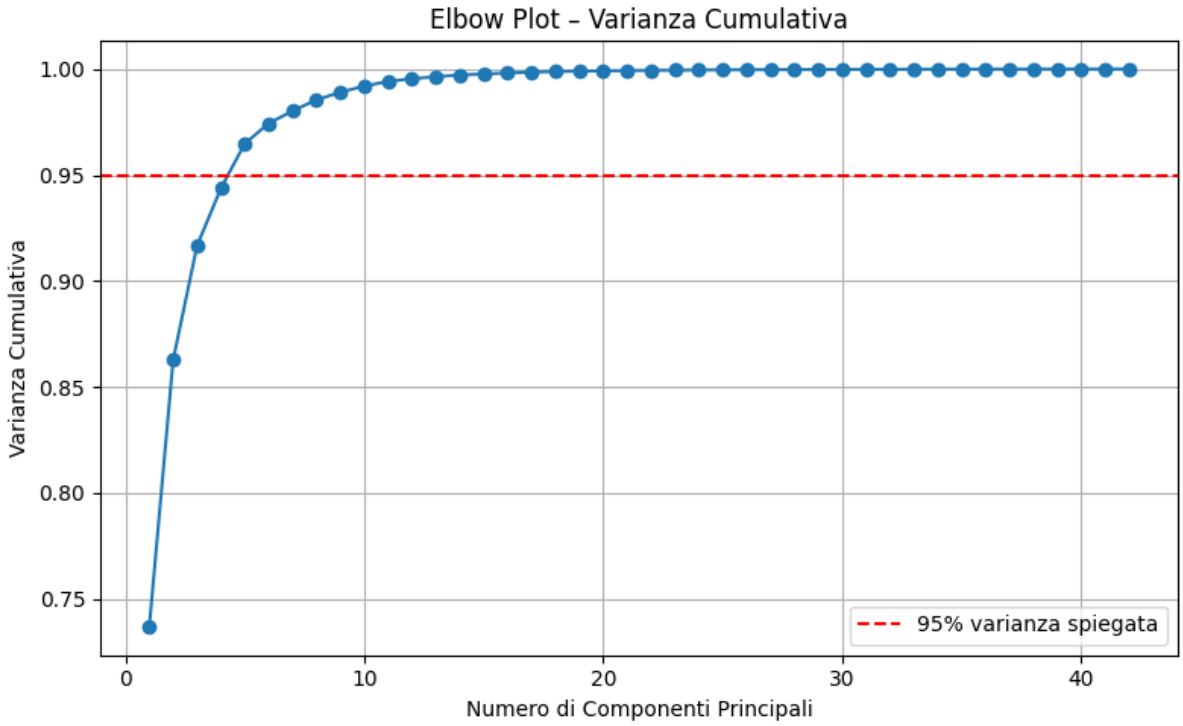


Figura 4.4 - Triticum durum e triticum monococcum - Elbow Plot relativo alla varianza cumulativa spiegata dalle componenti principali e indicazione della soglia al 95%.

Gli score plot bidimensionali ottenuti dall'analisi PCA sono rappresentati graficamente in Figura 4.5.

L'esame delle combinazioni tra le componenti principali evidenzia una discreta capacità di separazione tra le varietà di frumento analizzate, pur in presenza di una parziale sovrapposizione tra alcuni gruppi.

Nel piano PC1 vs PC2 si osserva una tendenza alla separazione della varietà Norberto (ciano) rispetto alle altre, con una distribuzione prevalente su valori positivi di PC1 e negativi di PC2. Anche Hammurabi (rosso) mostra una parziale distinzione, posizionandosi in zone differenti rispetto a Minosse (verde) e Creso (giallo).

Nei piani PC1 vs PC3 e PC1 vs PC4 la separazione della varietà Norberto appare ulteriormente rafforzata, con una collocazione stabile in regioni distinte rispetto agli altri gruppi. Le varietà Achille (blu) e Telemaco (viola) mostrano invece una maggiore sovrapposizione, suggerendo una composizione metabolica più simile tra loro.

L'analisi dei piani PC2 vs PC3 e PC2 vs PC4 mette in evidenza una dispersione interna più marcata, indicando che le componenti PC2, PC3 e PC4 descrivono principalmente la variabilità intragruppo piuttosto che differenze nette tra varietà differenti.

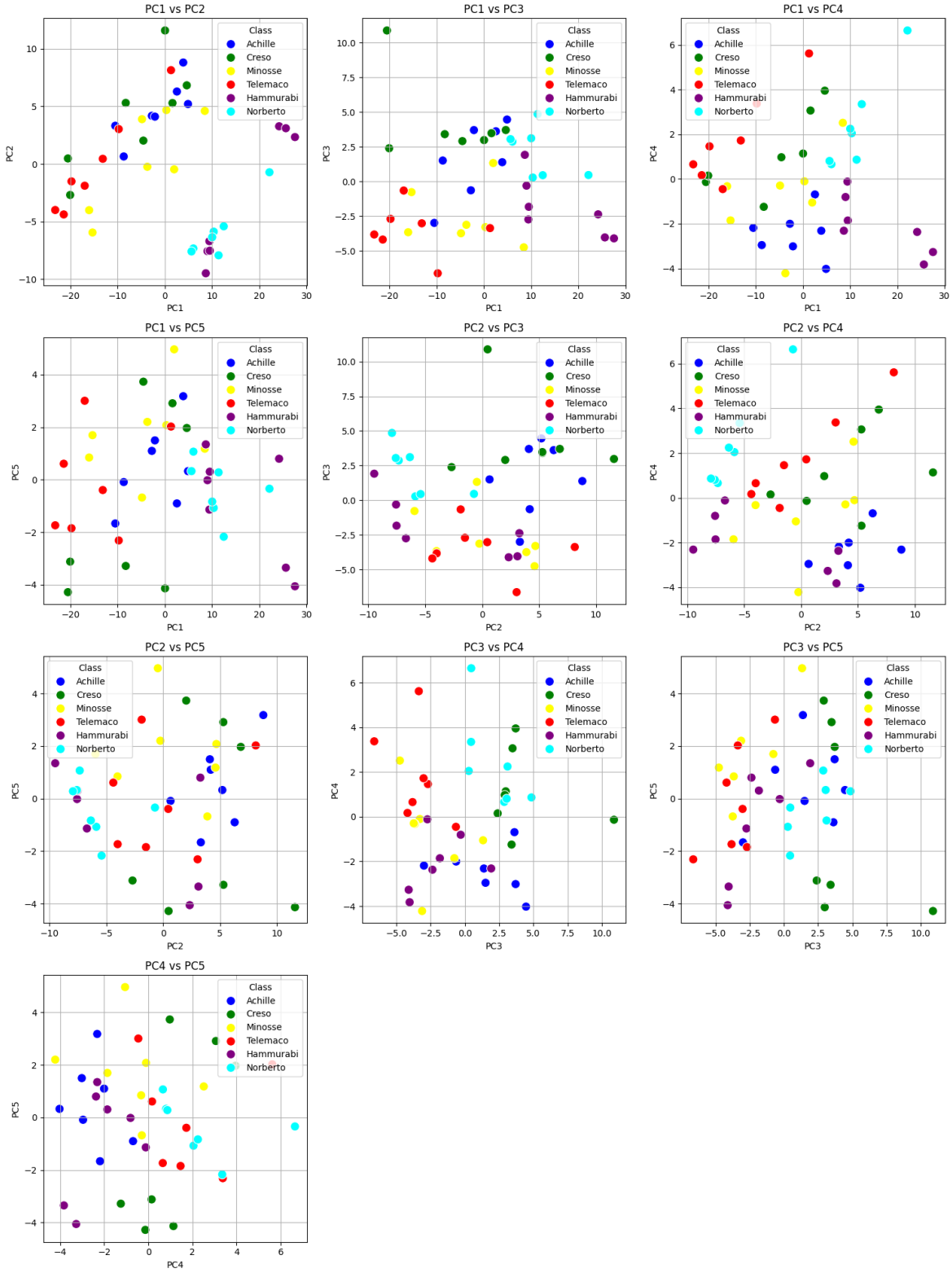


Figura 4.5 - Triticum durum e triticum monococcum - Distribuzione delle osservazioni nello spazio bidimensionale considerando la combinazione delle prime 5 PC.

Nei piani PC2 vs PC5 e PC3 vs PC5, varietà come Hammurabi e Norberto continuano a mantenere una parziale separazione rispetto agli altri gruppi, mentre Minosse, Achille e Creso risultano maggiormente sovrapposti, confermando una distanza metabolica inferiore tra queste varietà.

Infine, il piano PC4 vs PC5 mostra una capacità discriminativa piuttosto limitata, con un'ampia sovrapposizione tra tutte le varietà, suggerendo che le sole componenti PC4 e PC5 non risultano sufficientemente informative per garantire una chiara distinzione tra i gruppi analizzati.

Addestramento dei modelli di machine learning supervisionato

I risultati riportati nella Tabella 4.3 illustrano le performance predittive ottenute durante la fase di addestramento dei modelli di classificazione. L'analisi è stata condotta mediante validazione incrociata stratificata a 5 fold, in presenza di rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza dei modelli rispetto a una variabilità simulata. Per ciascun algoritmo sono riportate l'accuratezza media, la deviazione standard e gli iperparametri ottimali individuati tramite ottimizzazione con GridSearchCV.

Tabella 4.3 - Triticum durum e triticum monococcum - Prestazioni predittive dei modelli di classificazione durante l'addestramento con dati affetti da rumore gaussiano.

Modello	Accuracy (mean)	Accuracy (std)	Parametri ottimali
Random Forest	0.733	0.100	{'max_depth': None, 'n_estimators': 100}
SVM	0.900	0.146	{'C': 0.1, 'kernel': 'linear'}
XGBoost	0.689	0.118	{'learning_rate': 0.1, 'max_depth': 3..
k-NN	0.592	0.076	{'metric': 'manhattan', 'n_neighbors': 3}
Naive Bayes	0.664	0.059	default
Decision Tree	0.664	0.099	{'max_depth': None}
Logistic Regression	0.978	0.044	{'C': 10}

Nel complesso, le performance risultano piuttosto eterogenee, con valori di accuratezza compresi tra 0.592 (k-NN) e 0.978 (Logistic Regression). Questa variabilità può essere ricondotta sia alle differenze intrinseche tra gli algoritmi in termini di capacità discriminativa e sensibilità al rumore, sia alla perturbazione artificiale introdotta nei dati mediante l'aggiunta di rumore gaussiano. Trattandosi di una fase intermedia del workflow, i risultati ottenuti offrono un'indicazione preliminare sulla stabilità e sull'affidabilità dei modelli. Le reali capacità di generalizzazione verranno successivamente valutate attraverso la classificazione su un set esterno di campioni non affetti da rumore, con l'obiettivo di stimarne l'effettiva efficacia in condizioni operative.

Nei grafici seguenti sono riportate le metriche di performance dei modelli ottenute durante la fase di validazione su dati non precedentemente visti, affetti da rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza predittiva degli algoritmi in condizioni di lieve perturbazione.

La Figura 4.17 presenta un confronto diretto tra i modelli in termini di accuratezza media, evidenziando la percentuale complessiva di classificazioni corrette rispetto al totale delle osservazioni, mentre la Figura 4.18 riporta i valori di F1-score macro, metrica particolarmente utile per valutare la capacità del modello di bilanciare precisione e recall nelle classificazioni multi-classe. La Figura 4.19, infine, rappresenta la relazione tra precisione e recall macro, e permette di valutare l'equilibrio tra la capacità del modello di non produrre falsi positivi (precisione) e quella di rilevare correttamente le classi positive (recall).

L'analisi congiunta di queste metriche evidenzia una sostanziale coerenza con quanto osservato in fase di addestramento, confermando un buon livello di robustezza predittiva da parte della maggior parte dei modelli anche in presenza di una perturbazione moderata dei dati. In particolare, Logistic Regression si conferma il modello più performante, mantenendo elevati valori di accuratezza (vicina a 0.98), F1-score macro e un ottimo bilanciamento tra precisione e recall, a testimonianza di una spiccata capacità generalizzante.

Anche SVM e Random Forest mostrano performance solide e ben distribuite tra le varie metriche, con un buon compromesso tra sensibilità e specificità. Al contrario, algoritmi come k-NN, Naive Bayes e XGBoost, evidenziano una maggiore sensibilità al rumore, con una lieve riduzione dei valori di accuratezza e F1-score, nonché una collocazione meno favorevole nel grafico di precisione vs recall.

Questi risultati suggeriscono che, nel caso l'efficienza dei modelli possa essere fortemente influenzata sia dalla natura della matrice biologica sia dalla complessità intrinseca delle classi da discriminare. La fase successiva di validazione esterna, su campioni non affetti da rumore, sarà fondamentale per confermare la reale generalizzabilità e trasferibilità dei modelli sviluppati

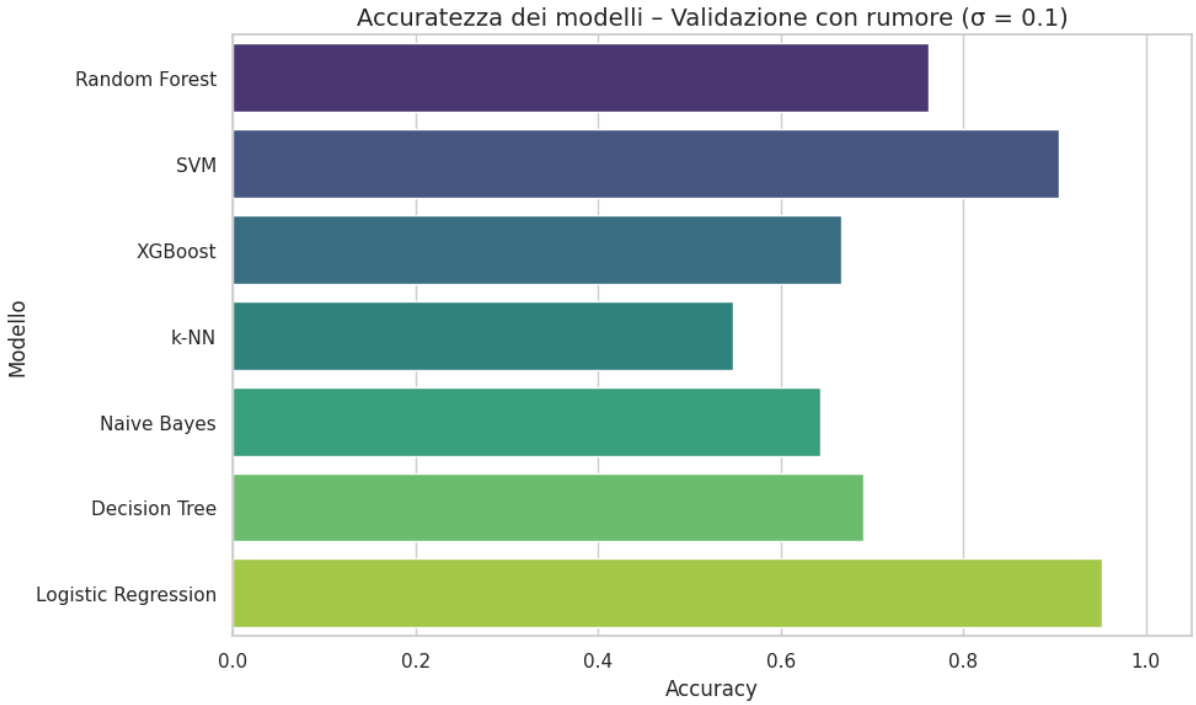


Figura 4.6 - Triticum durum e triticum monococcum - Accuracy per ciascun modello durante predizione su dati non visti.

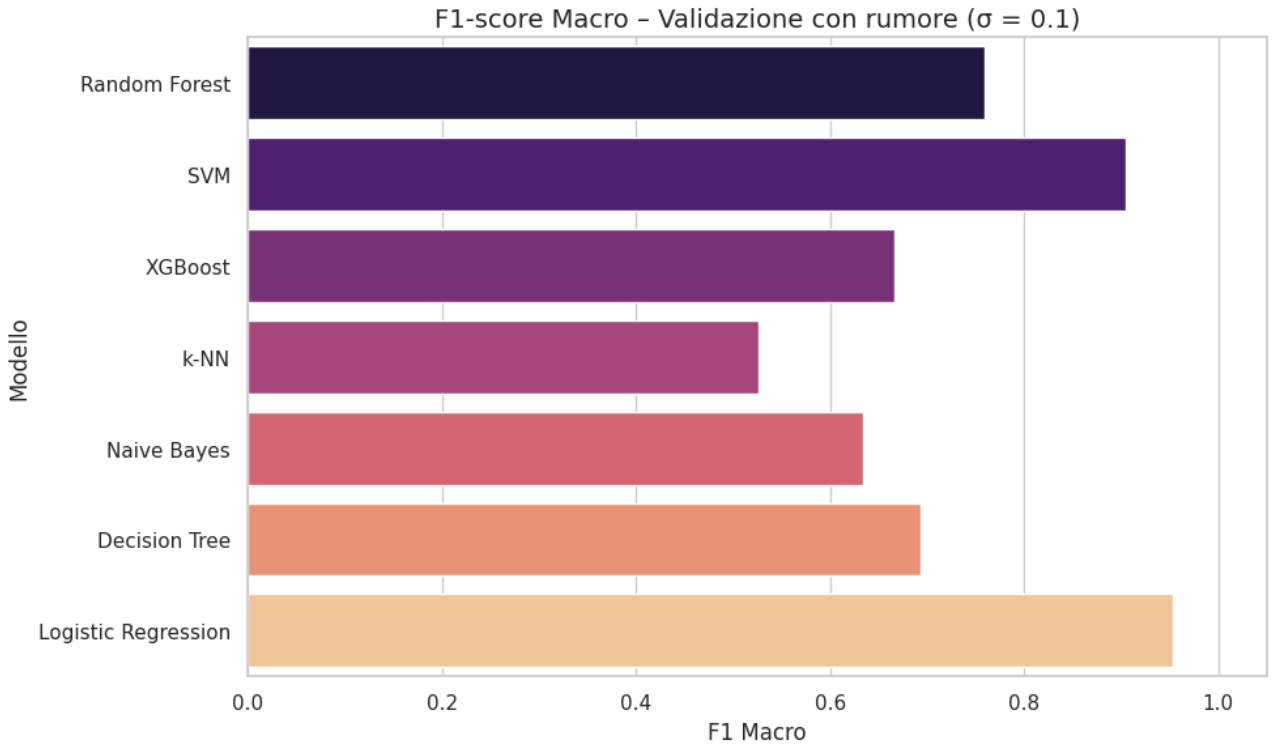


Figura 4.7 - Triticum durum e triticum monococcum - F1-score Macro per ciascun modello durante predizione su dati non visti.

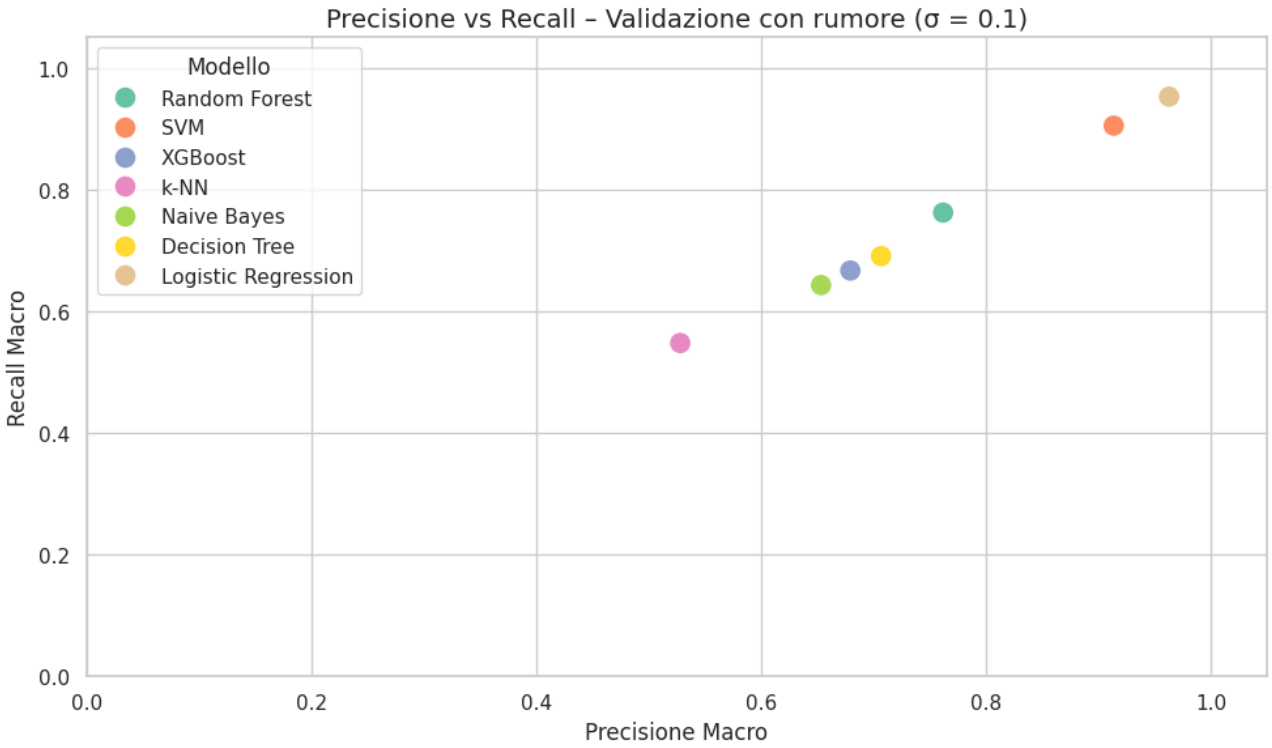


Figura 4.8 - Triticum durum e triticum monococcum - Precisione vs Recall macro per ciascun modello durante predizione su dati non visti.

Classificazione finale e selezione delle feature più rilevanti

I risultati riportati nella Tabella 4.4 evidenziano le prime 20 variabili (buckets NMR) che hanno mostrato la maggiore rilevanza durante il processo di classificazione condotto dai diversi algoritmi di machine learning, valutata mediante il criterio della *Mean Decrease Accuracy* (MDA). L'analisi si riferisce alla fase di test dei modelli eseguita su dati affetti da rumore gaussiano di intensità moderata ($\sigma = 0.1$), al fine di simulare condizioni realistiche di variabilità sperimentale.

Le variabili sono elencate in ordine decrescente rispetto al valore massimo di importanza MDA registrato, e sono accompagnate dagli intervalli spettrali corrispondenti (espressi in ppm) e dal numero di modelli predittivi nei quali sono risultate significative per la classificazione.

Tabella 4.4 – Triticum durum e triticum monococcum - Principali variabili (buckets NMR) selezionate tramite MDA durante il processo di classificazione.

Num	Variabile	PPM		Totale modelli	Importanza MDA
1	Var154	3.26	±0.02	7	2.362
2	Var153	3.30	±0.02	7	2.309
3	Var204	1.26	±0.02	7	1.846
4	Var141	3.78	±0.02	7	1.761
5	Var142	3.74	±0.02	7	1.751
6	Var149	3.46	±0.02	7	1.733
7	Var146	3.58	±0.02	7	1.721
8	Var143	3.70	±0.02	7	1.621
9	Var139	3.86	±0.02	6	1.527
10	Var137	3.94	±0.02	5	1.356
11	Var214	0.86	±0.02	6	1.319
12	Var138	3.90	±0.02	5	1.284
13	Var152	3.34	±0.02	7	1.269
14	Var151	3.38	±0.02	7	1.264
15	Var215	0.82	±0.02	4	1.220
16	Var216	0.78	±0.02	4	1.198
17	Var105	5.42	±0.02	4	1.181
18	Var219	0.66	±0.02	5	1.171
19	Var226	0.38	±0.02	3	1.137
20	Var136	3.98	±0.02	4	1.136

Le prime 20 variabili selezionate mostrano valori di MDA compresi tra 1.136% e 2.362%, delineando un pattern di importanza ben strutturato, con un nucleo di variabili ad alta rilevanza e un contributo progressivamente decrescente, suggerendo la presenza di alcuni segnali particolarmente incisivi nella discriminazione tra le classi.

Spiccano in particolare Var154 (3.26 ppm) e Var153 (3.30 ppm), entrambe selezionate da tutti e sette i modelli, con valori MDA rispettivamente pari a 2.362% e 2.309%, i più alti dell'intero set. La loro costante individuazione da parte di tutti gli algoritmi e l'elevata importanza associata confermano un ruolo centrale e sistematico nella classificazione.

Lo stesso comportamento si osserva per un ampio gruppo di variabili localizzate tra 3.26 e 3.74 ppm, tra cui Var204 (1.26 ppm), Var141 (3.78 ppm), Var142 (3.74 ppm), Var149 (3.46 ppm), Var146 (3.58 ppm), Var143 (3.70 ppm), e Var152–151 (3.34–3.38 ppm): tutte selezionate da tutti e sette i modelli e con MDA superiori a 1.2%, a indicare una zona spettrale fortemente rappresentata e discriminante.

Anche variabili con valori MDA leggermente inferiori, come Var139 (3.86 ppm, 1.527%) o Var214 (0.86 ppm, 1.319%), mostrano una buona stabilità (selezionate da 5 o 6 modelli), confermando la loro rilevanza, seppur con minore trasversalità.

Al contrario, le ultime variabili in tabella – come Var226 (0.38 ppm) e Var136 (3.98 ppm) – pur presentando MDA superiori all'1.1%, sono state selezionate da soli 3 o 4 modelli, suggerendo un contributo più limitato o meno generalizzabile, probabilmente legato a specificità algoritmiche o a dinamiche locali nei dati.

Validazione mediante set di dati esterno

Nella tabella seguente è riportata la performance dei diversi modelli in termini di accuratezza nella predizione su un set di dati esterno, privo di rumore gaussiano, ottenuto come descritto nel paragrafo 3.2.2.1.

In Figura 4.9 viene presentata, a titolo esemplificativo, la matrice di confusione relativa al modello Random Forest ottenuta nella fase di predizione sul set esterno.

Tabella 4.5 - Triticum durum e triticum monococcum - Accuratezza, F1-score macro e F1-score weighted dei modelli di classificazione applicati al set esterno di dati.

Modello	Accuracy	F1_macro	F1_weighted
Random Forest	1.000	1.000	1.000
SVM	1.000	1.000	1.000
XGBoost	0.952	0.952	0.952
k-NN	0.881	0.879	0.879
Naive Bayes	0.929	0.928	0.928
Decision Tree	0.976	0.976	0.976
Logistic Regression	1.000	1.000	1.000

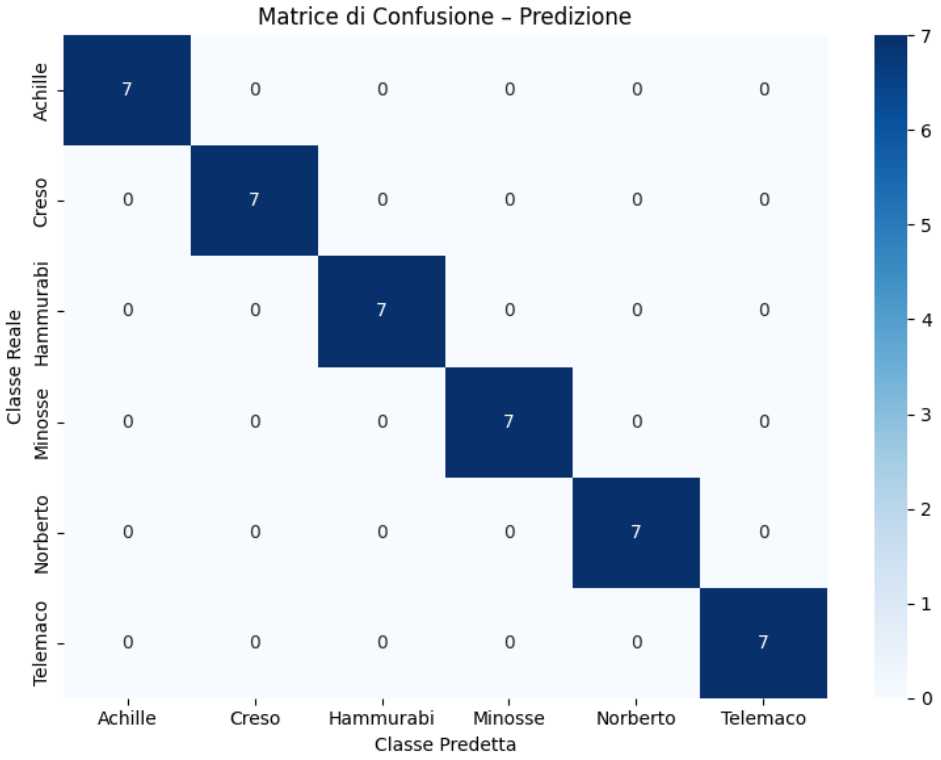


Figura 4.9 - Triticum durum e triticum monococcum - Matrice di confusione ottenuta con il modello Random Forest durante classificazione con set esterno di dati..

I risultati della predizione sui dati esterni evidenziano prestazioni particolarmente elevate da parte dei modelli valutati. In particolare, Random Forest, SVM e Logistic Regression hanno raggiunto un'accuratezza pari al 1.00, confermata anche dagli F1-score macro e weighted, suggerendo una perfetta capacità di generalizzazione sui campioni esterni, senza incorrere in errori di classificazione.

Il modello Decision Tree ha mostrato anch'esso un'ottima performance, con un'accuratezza pari a 0.976, evidenziando soltanto minime discrepanze rispetto alla predizione ideale. XGBoost si è attestato su un valore di accuratezza pari a 0.952, mantenendo comunque un eccellente bilanciamento tra precisione e richiamo, come indicato dagli F1-score macro e weighted identici.

I modelli k-NN e Naive Bayes, pur presentando buoni livelli di accuratezza (rispettivamente 0.881 e 0.929), hanno mostrato prestazioni inferiori rispetto agli altri algoritmi, confermando quanto già osservato durante la fase di validazione interna. In particolare, la minore accuratezza del k-NN potrebbe essere imputabile a una maggiore sensibilità alle variazioni locali della distribuzione dei dati nello spazio delle componenti principali.

L’analisi della matrice di confusione per Random Forest evidenzia un risultato perfetto in termini di classificazione: tutti i campioni (7 per ciascuna classe) sono stati correttamente assegnati alla propria categoria di appartenenza, senza errori di classificazione. Questo comportamento indica che il modello ha generalizzato in modo estremamente efficace durante la precedente fase di addestramento (su dati affetti da rumore gaussiano), confermando un’elevata robustezza e una forte capacità discriminante su un set esterno non visto.

Discussione finale

L’approccio automatizzato basato sull’integrazione tra PCA e modelli di machine learning supervisionato ha mostrato una discreta efficacia nell’analisi e classificazione della matrice Triticum durum e monococcum.

I modelli predittivi addestrati sulle componenti latenti, in presenza di rumore gaussiano (σ = 0.1), hanno evidenziato performance eterogenee ma complessivamente solide. Logistic Regression e SVM si sono dimostrati i più efficaci durante la validazione interna, seguiti da Random Forest, mentre modelli come k-NN e Naive Bayes hanno mostrato maggiore variabilità e sensibilità alla perturbazione dei dati. Le differenze osservate riflettono la diversa capacità dei modelli di adattarsi a pattern metabolici parzialmente sovrapposti e a una complessità classificatoria che richiede robustezza nei confronti di rumore e ridondanza informativa.

La successiva selezione delle variabili più rilevanti tramite MDA ha confermato la presenza di segnali spettrali altamente informativi, con valori di importanza compresi tra 1.136% e 2.362%. In particolare, le variabili comprese tra 3.26 e 3.74 ppm sono state selezionate da tutti i modelli e presentano valori MDA costantemente elevati, indicando un chiaro contributo discriminante. La coerenza tra le variabili selezionate e il numero di modelli che le hanno identificate come rilevanti sottolinea la stabilità di questi segnali nella caratterizzazione varietale.

Infine, la validazione su un set esterno di dati, non affetti da rumore, ha rappresentato un passaggio cruciale per la verifica della capacità di generalizzazione dei modelli. I risultati ottenuti sono stati ottimi per gran parte dei modelli: in particolare, Logistic Regression, SVM e Random Forest hanno raggiunto un'accuratezza del 100%, confermata anche dagli F1-score macro e weighted, a dimostrazione di un apprendimento robusto, ben generalizzabile e privo di overfitting. La matrice di confusione associata a Random Forest mostra una classificazione perfetta per ciascuna delle sei classi analizzate.

Nel complesso, l’intero workflow AI-driven si è dimostrato altamente performante, riproducibile e adattabile a dati complessi come quelli derivati da spettroscopia NMR. La

combinazione di riduzione dimensionale, addestramento supervisionato e selezione delle feature ha permesso non solo di raggiungere un'elevata accuratezza nella classificazione, ma anche di individuare le regioni spettrali più informative per la discriminazione tra le classi.

4.1.4 Confronto tra approccio classico e automatizzato

Il confronto tra l'approccio chemiometrico convenzionale e quello automatizzato AI-driven evidenzia vantaggi complementari e alcune differenze sostanziali nelle strategie di analisi e nelle capacità discriminative.

Dal punto di vista esplorativo, l'analisi PCA condotta con metodo classico ha permesso di identificare chiaramente le principali fonti di variabilità, con una separazione netta delle varietà appartenenti a specie diverse (T. durum vs. T. monococcum) e una buona caratterizzazione delle varietà moderne di frumento duro. La selezione delle variabili discriminanti, basata su contributi percentuali e conferma statistica (ANOVA + test di Tukey), ha permesso l'assegnazione putativa di numerosi segnali a metaboliti noti, fornendo un'interpretazione biochimica diretta delle differenze tra classi.

L'approccio automatizzato, basato sull'integrazione tra PCA, modelli supervisionati e analisi MDA, ha mostrato un'ottima efficacia nella fase predittiva. I modelli Logistic Regression, SVM e Random Forest hanno raggiunto, nella validazione esterna, un'accuratezza perfetta (1.000), con valori di F1-score macro e weighted altrettanto elevati. Tale risultato, confermato dalla matrice di confusione di Random Forest, suggerisce una capacità di generalizzazione molto elevata e l'assenza di overfitting. Rispetto al metodo classico, il workflow automatizzato ha inoltre evidenziato un miglior controllo del rumore e una selezione delle feature più stabile, supportata dal fatto che molte delle variabili con MDA elevata sono state selezionate da tutti e sette i modelli. Un aspetto particolarmente rilevante emerso dal confronto riguarda la coerenza tra le variabili selezionate nei due approcci. In particolare, l'analisi automatizzata ha identificato come altamente rilevanti numerosi segnali localizzati nella regione spettrale δ 3.26–3.74 ppm, tra cui le variabili Var154, Var153, Var152, Var151, Var149, Var146, Var143, Var142 e Var141, tutte selezionate da 6 o 7 modelli su 7 e caratterizzate da valori di MDA superiori a 1.2%. Questo cluster di segnali corrisponde esattamente all'area in cui anche l'approccio classico aveva evidenziato il contributo più marcato delle componenti PC2 e PC3, identificando metaboliti come glucosio (Var146, Var148, Var149, Var150), fruttosio (Var145), serina (Var144) e colina/etanolamina (Var152). Questa sovrapposizione tra variabili significative indica una notevole convergenza tra i due metodi, nonostante le differenze nei criteri di selezione. Le stesse variabili emergono sia in base all'analisi statistica classica (ANOVA +

Tukey sui contributi PCA), sia tramite ranking automatizzato dell'importanza predittiva (MDA), rafforzando l'affidabilità delle informazioni spettrali identificate. Inoltre, entrambe le strategie evidenziano il ruolo cruciale di metaboliti zuccherini e amminoacidici nella discriminazione tra le varietà di Triticum monococcum e durum, confermando che le differenze varietali sono riflettute principalmente nella composizione glucidica e nel metabolismo primario azotato. Questa convergenza qualitativa e quantitativa tra i due approcci rappresenta un elemento di validazione incrociata, che rafforza la solidità del protocollo proposto.

4.1.5 Esempio di serializzazione e generazione hash per registrazione del fingerprint metabolico su blockchain

Di seguito viene mostrato un esempio di file serializzato in formato JSON ottenuto a partire dal segnale NMR grezzo (FID) acquisito da un campione rappresentativo della matrice Triticum durum (codice TD_Achille6Farina_pD2O), attraverso spettroscopia 1H HR-MAS.


```
{
  "sample_id": "TD_Achille6Farina_pD2O_23_07_2021",
  "matrix": "Triticum durum",
  "instrument": "av400",
  "operator": "Cangemi",
  "date": "2021-07-23T14:47:22",
  "temperature": 298.3419,
  "pulse_program": "zgpr",
  "fid_real": [
    0.0,
    0.0,
…
    137.0
  ],
  "fid_imag": [
    0.0,
    0.0,
…
    -308.0
  ],
  "is_domain": "time",
  "timestamp": "2025-06-10T06:30:10.734296"
}
```


Figura 4.10 – Triticum durum e triticum monococcum - Esempio di serializzazione da FID Bruker

Il file JSON include la parte reale e quella immaginaria del FID e un insieme di metadati contestuali: strumento, operatore, data, temperatura, sequenza di impulso e timestamp.
Il contenuto del file JSON, sottoposto a funzione di hash SHA256, ha fornito il seguente identificativo digitale:

a4dac62618c3d1dbe78508198a53c6b474ff0426f3302aa418c2b12f426550d6

Figura 4.11 - Triticum durum e triticum monococcum - Esempio di generazione hash da FID Bruker

L’hash così ottenuto può fungere da identificativo digitale non modificabile, utile per ancorare il dato analitico a uno smart contract distribuito.
Questo hash rappresenta una "firma digitale" univoca del file, che ne garantisce l’integrità e l’autenticità. Una volta registrato su una rete blockchain (es. Hyperledger Fabric), tale impronta permette di verificare, in qualsiasi momento, che il file non sia stato modificato e che l’acquisizione sia avvenuta in condizioni sperimentali note e tracciabili.

4.2 Triticum durum - Varietà “Senatore Cappelli”

4.2.1 Descrizione dei profili spettrali

Nella Figura 4.12 sono riportati gli spettri 1H HR-MAS NMR relativi ai campioni di grano duro varietà 'Senatore Cappelli', coltivati in differenti aree geografiche italiane. Analogamente a quanto osservato per le altre varietà di frumento analizzate, anche i campioni di Triticum durum varietà 'Senatore Cappelli' mostrano un profilo spettrale coerente con la composizione metabolica tipica della matrice. Le principali aree di segnale – alifatica (δ 0.8–2.5 ppm), zuccherina (δ 3.0–5.5 ppm) e aromatica (δ 6.0–9.5 ppm) – risultano ben rappresentate, riflettendo la presenza di amminoacidi ramificati, zuccheri solubili, fruttani e composti fenolici, come già descritto nel paragrafo precedente. Tuttavia, nonostante la similarità del profilo globale, si osservano variazioni interessanti tra i campioni provenienti da diverse aree di coltivazione. In particolare, nella regione zuccherina (δ 3.0–5.5 ppm) si notano differenze di intensità tra i segnali più abbondanti, verosimilmente riconducibili a variazioni nei livelli di glucosio, fruttosio e/o saccarosio. Ad esempio, il profilo del campione 'Cavaliere' evidenzia una maggiore intensità relativa in quest’area rispetto agli altri, suggerendo una differente composizione carboidratica, probabilmente influenzata da pratiche agronomiche o condizioni pedoclimatiche specifiche. Anche nella regione alifatica (δ 0.8–2.5 ppm) si rilevano variazioni nei segnali associabili ad acidi grassi o metaboliti lipidici, con bande più marcate in campioni come 'Ciacci' e 'De Ioanni', il che potrebbe riflettere differenze nel grado di maturazione della cariosside o nella composizione proteica. Infine, nella regione aromatica (δ 6.0–9.5 ppm), pur essendo i segnali generalmente meno intensi, si osserva una modulazione dell’intensità nei campioni 'Colucci' e 'Crea1C', potenzialmente attribuibile a variazioni nella concentrazione di composti fenolici, noti per il loro ruolo nella risposta agli stress ambientali e nella modulazione della qualità del frumento.

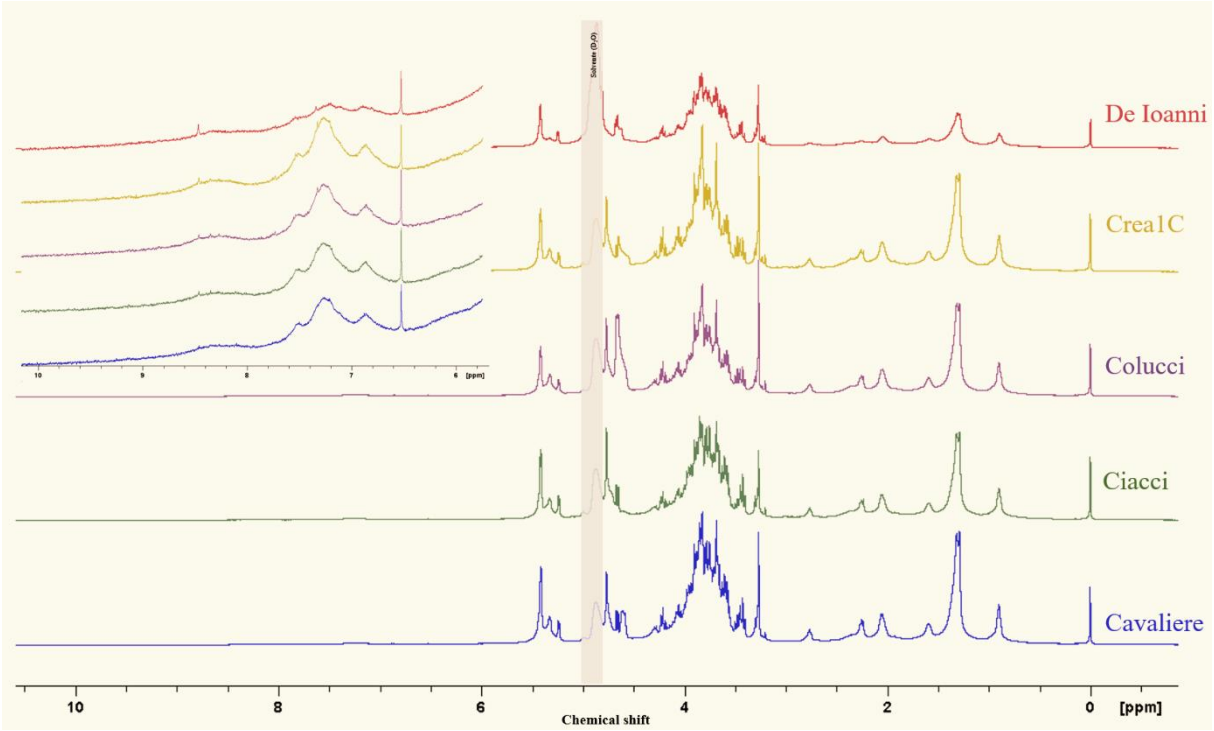


Figura 4.12 - Triticum durum (varietà Senatore Cappelli) - Spettri 1H-NMR.

4.2.2 Risultati dell'analisi chemiometrica convenzionale

PCA

La

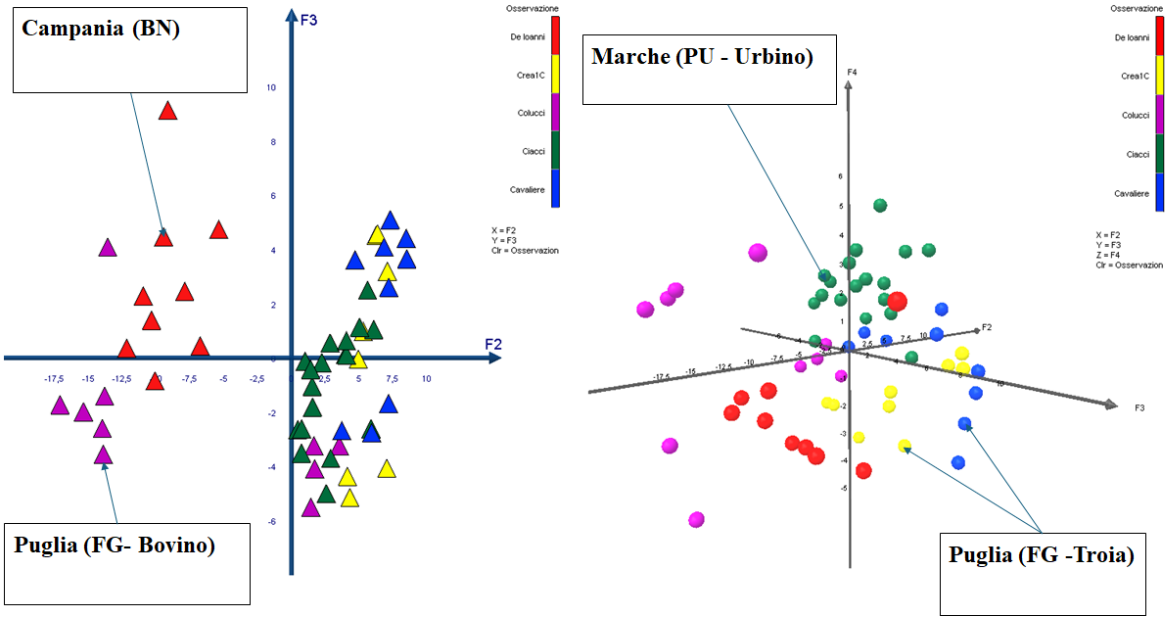


Figura 4.13 mostra due score plot rappresentativi ottenuti mediante analisi PCA in XLSTAT, relativi ai campioni di grano duro varietà 'Senatore Cappelli'. Nello score plot PC2–PC3 è evidente una separazione geografica marcata, con i campioni provenienti dalla Campania (BN)

e dalla Puglia (FG - Bovino) distribuiti in regioni distinte lungo le componenti principali. Tale separazione suggerisce la presenza di differenze metaboliche associate all'area di coltivazione, verosimilmente legate a fattori ambientali, agronomici o pedoclimatici. L'analisi tridimensionale nel piano definito dalle componenti PC2, PC3 e PC4 conferma e approfondisce questa osservazione, rivelando una distinzione ancora più marcata nella distribuzione spaziale delle osservazioni. In particolare, i campioni provenienti dalle Marche (PU - Urbino) tendono a raggrupparsi in una regione separata rispetto a quelli della Puglia (FG - Troia), evidenziando una chiara organizzazione dei dati in funzione dell'origine geografica. Nel dettaglio, i campioni dell'azienda De Ioanni (Campania, BN) si concentrano nel quadrante superiore sinistro del piano PC2–PC3, separandosi nettamente dagli altri. I campioni dell'azienda Colucci (Puglia, FG - Bovino) presentano a loro volta una distribuzione compatta e distinta, indicativa di un'elevata coerenza metabolica interna. Al contrario, le aziende CreaC1 e Cavaliere, pur appartenendo alla stessa area geografica (Troia, FG), mostrano una certa variabilità lungo PC3, suggerendo microdifferenze pedologiche o pratiche agronomiche differenti (Radaelli et al., 2024). I campioni dell'azienda Ciacci (Marche, PU - Urbino) si distribuiscono invece in posizione più centrale, suggerendo un profilo metabolico intermedio, probabilmente influenzato da condizioni climatiche più temperate.

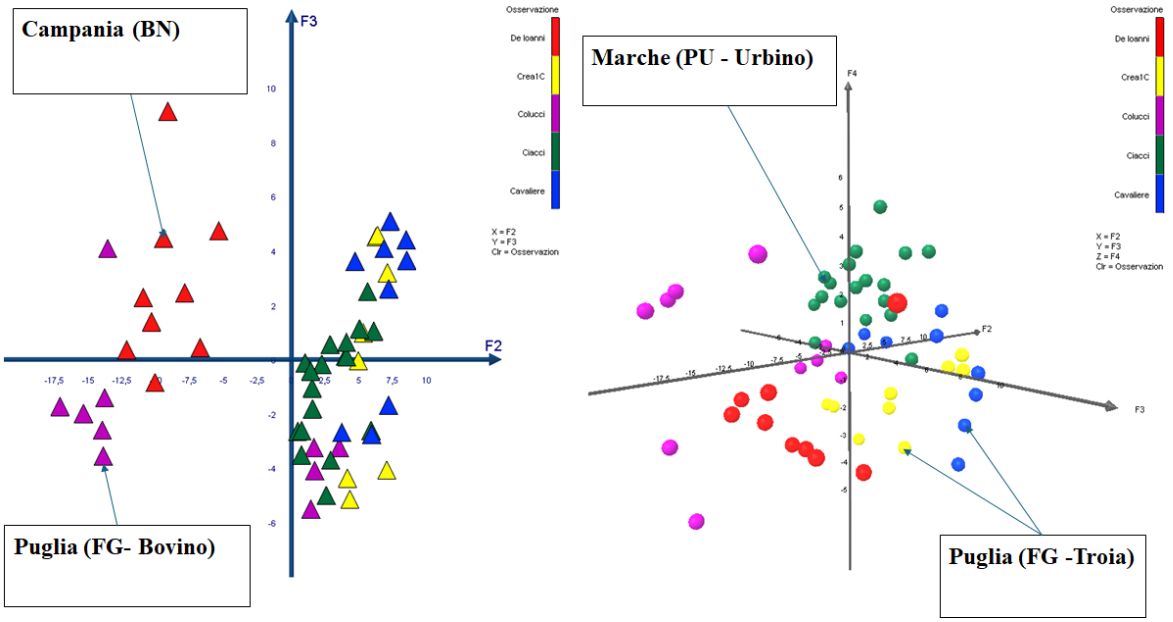


Figura 4.13 - Triticum durum (varietà Senatore Cappelli) - Score Plot da PCA.

Selezione delle variabili significative

La Tabella 4.6 riporta, per ciascuna delle prime cinque componenti principali, l'elenco delle variabili che risultano significativamente differenti tra almeno due gruppi ($p < 0.05$), sulla base dell'analisi della varianza univariata (ANOVA). È opportuno precisare che la significatività statistica indicata non implica che ciascuna variabile discrimini tutte le classi tra loro, bensì che, per almeno un confronto tra coppie di gruppi, è stata rilevata una differenza significativa.

Per una maggiore chiarezza interpretativa, è stata successivamente condotta un'analisi post-hoc (test di Tukey HSD) al fine di individuare con precisione quali gruppi differiscono tra loro per ciascuna variabile significativa. Nella presente tabella è tuttavia riportato esclusivamente l'elenco complessivo delle variabili che hanno mostrato una significatività globale (ANOVA a $p < 0.05$), senza entrare nel dettaglio dei confronti specifici tra classi. Tale approccio è stato adottato con l'obiettivo di fornire una sintesi preliminare delle variabili maggiormente coinvolte nella separazione dei gruppi lungo ciascuna componente principale.

Tabella 4.6 - Triticum durum (varietà Senatore Cappelli) - Elenco delle variabili significative.

PC	Variabili significative ($p < 0.05$)
PC1	Var27, Var30, Var31, Var32, Var36, Var51, Var52, Var53, Var54, Var55, Var56, Var61, Var62, Var63, Var64, Var68, Var69, Var70, Var71, Var125, Var126
PC2	Var105, Var106, Var107, Var108, Var112, Var113, Var114, Var115, Var116, Var139, Var141, Var142, Var144, Var146, Var147, Var149, Var150, Var151, Var166, Var167, Var178, Var179, Var180, Var183, Var184, Var185, Var195, Var196, Var201, Var202, Var203, Var212, Var213
PC3	Var100, Var101, Var102, Var103, Var104, Var105, Var106, Var135, Var136, Var137, Var138, Var139, Var141, Var142, Var143, Var144, Var145, Var146, Var149, Var150, Var152, Var153
PC4	Var1, Var2, Var3, Var4, Var5, Var6, Var8, Var110, Var122, Var123, Var124, Var125, Var126, Var146, Var150, Var157, Var158, Var159, Var160, Var161, Var162, Var163, Var164, Var168, Var169, Var170, Var171, Var172, Var173, Var174
PC5	Var107, Var121, Var155, Var156, Var202, Var203, Var220, Var221

Assegnazione delle variabili discriminanti ai metaboliti

Nella Tabella 4.7 è riportata l'assegnazione delle principali variabili (bucket spettrali) a metaboliti putativi. L'integrazione tra le informazioni spettroscopiche ottenute dall'analisi 2D HSQC e i risultati statistici derivati dalla PCA, combinati con test ANOVA e post-hoc di Tukey, ha consentito di selezionare un gruppo di segnali che, in base alla loro posizione chimica e al contributo multivariato, risultano compatibili con metaboliti noti. Sebbene le assegnazioni proposte vadano considerate come putative. La distribuzione dei segnali lungo più componenti principali (PC) rafforza l'ipotesi che alcuni metaboliti – come zuccheri semplici, amminoacidi e composti azotati – giochino un ruolo centrale nei meccanismi di differenziazione tra campioni. Tuttavia, tali interpretazioni saranno oggetto di indagine supplementare, alla luce della complessità della matrice e della possibile sovrapposizione di segnali in regioni chiave dello spettro.

Tabella 4.7 - Triticum durum (varietà Senatore Cappelli) - Assegnazione putativa delle variabili ai metaboliti.

Buckets	PC	δ ^{1}H (ppm)	δ ^{13}C (ppm)	Metabolita
Var105	PC2, PC3	5,40	95,09	Glucosio (H1)
Var106	PC2, PC3	5,40	103,91	Saccarosio (H1)
Var110	PC4	5,22	94,95	Glucosio (H1)
Var136	PC3	4,17	63,58	Fruttosio
Var145	PC3	3,81	63,58	Serina
Var150	PC2, PC3, PC4	3,63	79,43	Glucosio (H2–H6)
Var155	PC5	3,41	72,44	Glucosio (H2–H6)
Var159	PC4	3,26	76,95	Glucosio (H2–H6)
Var159	PC4	3,25	56,12	Etanolamina / Colina
Var201	PC2	1,58	27,51	Alanina (CH3)

Tra i segnali più evidenti, quelli localizzati intorno a δ^{1}H 5.40 / δ^{13}C 95–104 ppm risultano compatibili con i protoni anomerici del glucosio e del saccarosio, due zuccheri semplici la cui presenza potrebbe variare tra i campioni esaminati. Tali segnali, visibili principalmente nelle PC2 e PC3, suggeriscono un potenziale contributo del metabolismo zuccherino nella spiegazione della variabilità osservata. All'interno della PC3, è stato inoltre individuato un segnale in δ^{1}H 4.17 / δ^{13}C 63.6 ppm, compatibile con il fruttosio, e un ulteriore picco (δ^{1}H 3.81 / δ^{13}C 63.6 ppm) potenzialmente associabile alla serina, entrambi metaboliti che potrebbero riflettere differenze nel contenuto zuccherino e amminoacidico tra le varietà o condizioni agronomiche studiate. Un altro segnale di interesse (δ^{1}H 3.25 / δ^{13}C 56.1 ppm), rilevato nella PC4, è attribuibile alla etanolamina o alla colina, suggerendo un possibile coinvolgimento di composti azotati nei meccanismi di differenziazione. Sempre nella stessa componente, la presenza di segnali laterali del glucosio potrebbe indicare la partecipazione di specifici adattamenti biochimici alla variabilità tra campioni. In PC5, è emerso un segnale secondario del glucosio (δ^{1}H 3.41 / δ^{13}C 72.4 ppm), mentre nella PC2 è stata riscontrata una risonanza compatibile con l'alanina (δ^{1}H 1.58 / δ^{13}C 27.5 ppm), un amminoacido spesso implicato in processi di stress e nel metabolismo intermedio. È infine interessante notare come alcune variabili, come ad esempio Var150 (associata al glucosio), compaiano simultaneamente in più componenti (PC2, PC3 e PC4), rafforzando l'ipotesi che alcuni segnali possano avere un ruolo trasversale nella discriminazione dei campioni. Pur trattandosi di assegnazioni putative, questa convergenza tra evidenza statistica e dati spettrali offre indicazioni coerenti con la letteratura, suggerendo un insieme di possibili biomarcatori di interesse per la caratterizzazione metabolomica del frumento.

4.2.3 Risultati dell'analisi chemiometrica AI-driven

Importazione, verifica e pre-processing dei dati

La Figura 4.14 mostra un estratto della matrice dei bucket spettrali importata nello script, rappresentando i dati prima e dopo la standardizzazione mediante trasformazione Z-score.
Tale standardizzazione realizza una distribuzione centrata attorno allo zero e priva di dominanze numeriche, condizione ottimale per l'applicazione di modelli di machine learning supervisionato. Questo passaggio garantisce che ogni variabile contribuisca in egual misura al modello, prevenendo distorsioni dovute alla diversa scala delle intensità.

Dati Grezzi (prime 5 righe):

	replica	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	...	Var222	Var223
0	1	1112830.89	1324066.01	1339688.55	1389248.56	1341592.97	1464328.01	1456848.22	1637402.83	1680482.99	...	17910586.87	16545923.13
1	2	1276087.70	1439154.94	1411905.12	1474453.26	1588295.41	1597683.42	1773251.72	1862884.21	1950009.20	...	18449581.84	17090757.03
2	3	997728.82	1053972.04	1036782.57	1151054.43	1176575.57	1275333.64	1348393.24	1307481.09	1329063.82	...	17543147.02	16102044.12
3	4	1001344.58	990463.88	983584.50	1023782.64	1030743.39	1124160.66	1252136.38	1265989.43	1260288.39	...	17098870.60	15796525.38
4	5	1232849.92	1349238.90	1372720.67	1415949.67	1480421.57	1524092.85	1591297.97	1741202.31	1736652.68	...	17435177.63	16148672.28

5 rows × 232 columns

Dati Standardizzati (prime 5 righe):

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224
0	0.023745	0.379707	0.323600	0.291912	0.048988	0.156882	0.048710	0.251741	0.208903	0.250797	...	0.528930	0.420101	0.406188	0.373894
1	0.391227	0.631296	0.484739	0.474441	0.527199	0.415652	0.653773	0.672327	0.682085	0.411896	...	0.687552	0.663482	0.672302	0.662459
2	-0.235343	-0.210729	-0.352280	-0.218355	-0.270884	-0.209853	-0.158690	-0.363655	-0.408052	-0.281017	...	0.303271	0.254184	0.189384	0.197445
3	-0.227204	-0.349560	-0.470981	-0.491001	-0.553567	-0.503197	-0.342764	-0.441048	-0.528794	-0.489638	...	0.107265	0.053572	0.040159	0.040683
4	0.293901	0.434736	0.397306	0.349112	0.318095	0.272853	0.305820	0.445356	0.307514	0.271942	...	0.195063	0.205431	0.212159	0.239947

5 rows × 230 columns

Figura 4.14 – Triticum durum (varietà Senatore Cappelli) - Estratto della matrice dei bucket spettrali prima e dopo la standardizzazione Z-score

Riduzione dimensionale mediante PCA

In analogia con quanto mostrato nel paragrafo 4.2.24.4.2, vengono presentati i risultati dell'analisi PCA condotta con l'impiego dello script. La Figura 4.15 riporta l'andamento della varianza cumulativa spiegata (Elbow Plot), evidenziando come la soglia del 95% venga raggiunta entro le prime cinque componenti principali. La combinazione dei criteri adottati — soglia del 95% della varianza spiegata, criterio di Kaiser e metodo del "Broken Stick" — ha condotto alla selezione delle prime cinque componenti principali, ritenute ottimali per l'addestramento e la predizione dei modelli di machine learning.

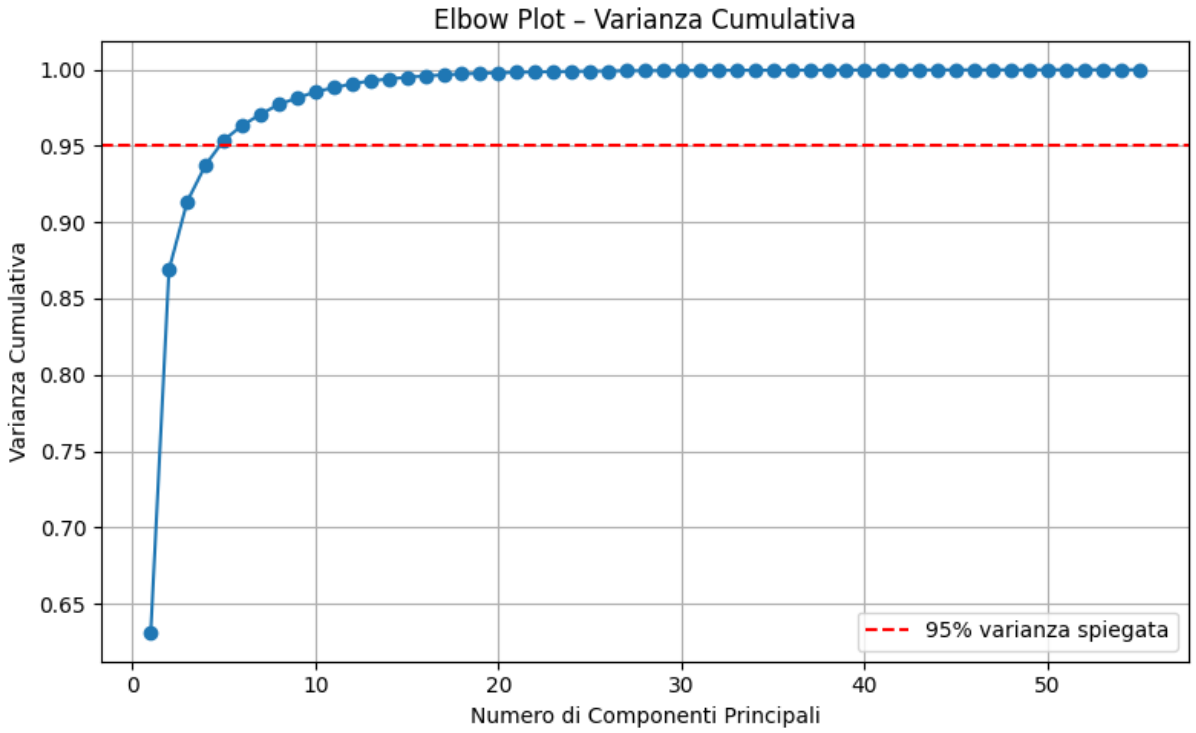


Figura 4.15 - Triticum durum (varietà Senatore Cappelli) - Elbow Plot relativo alla varianza cumulativa spiegata dalle componenti principali e indicazione della soglia al 95%.

Gli score plot bidimensionali ottenuti dall'analisi PCA automatizzata sono riportati in Figura 4.16.

L'esame delle combinazioni tra le componenti principali mostra una buona capacità di discriminazione tra i campioni provenienti da diverse aziende agricole, pur in presenza di alcune sovrapposizioni parziali.

Nel piano PC1 vs PC2 emerge chiaramente una separazione del gruppo De Ioanni (rosso), collocato in posizione distinta rispetto agli altri, con valori fortemente negativi lungo PC2. Anche Colucci (viola) mostra una distribuzione ben definita lungo PC1, mentre gli altri gruppi (Crea1C – giallo, Cavaliere – blu e Ciacci – verde) risultano in parte sovrapposti.

Nei piani PC1 vs PC3 e PC1 vs PC4, la separazione dei campioni De Ioanni rimane evidente, con un posizionamento stabile in aree isolate dello spazio multivariato. I campioni Colucci e Cavaliere si distribuiscono in zone intermedie, mentre Crea1C e Ciacci mostrano una maggiore eterogeneità interna.

Le combinazioni PC2 vs PC3 e PC2 vs PC4 mettono in luce una buona separazione di Colucci e De Ioanni, mentre Crea1C, Cavaliere e Ciacci tendono a occupare aree contigue, suggerendo un certo grado di similarità nei profili metabolici.

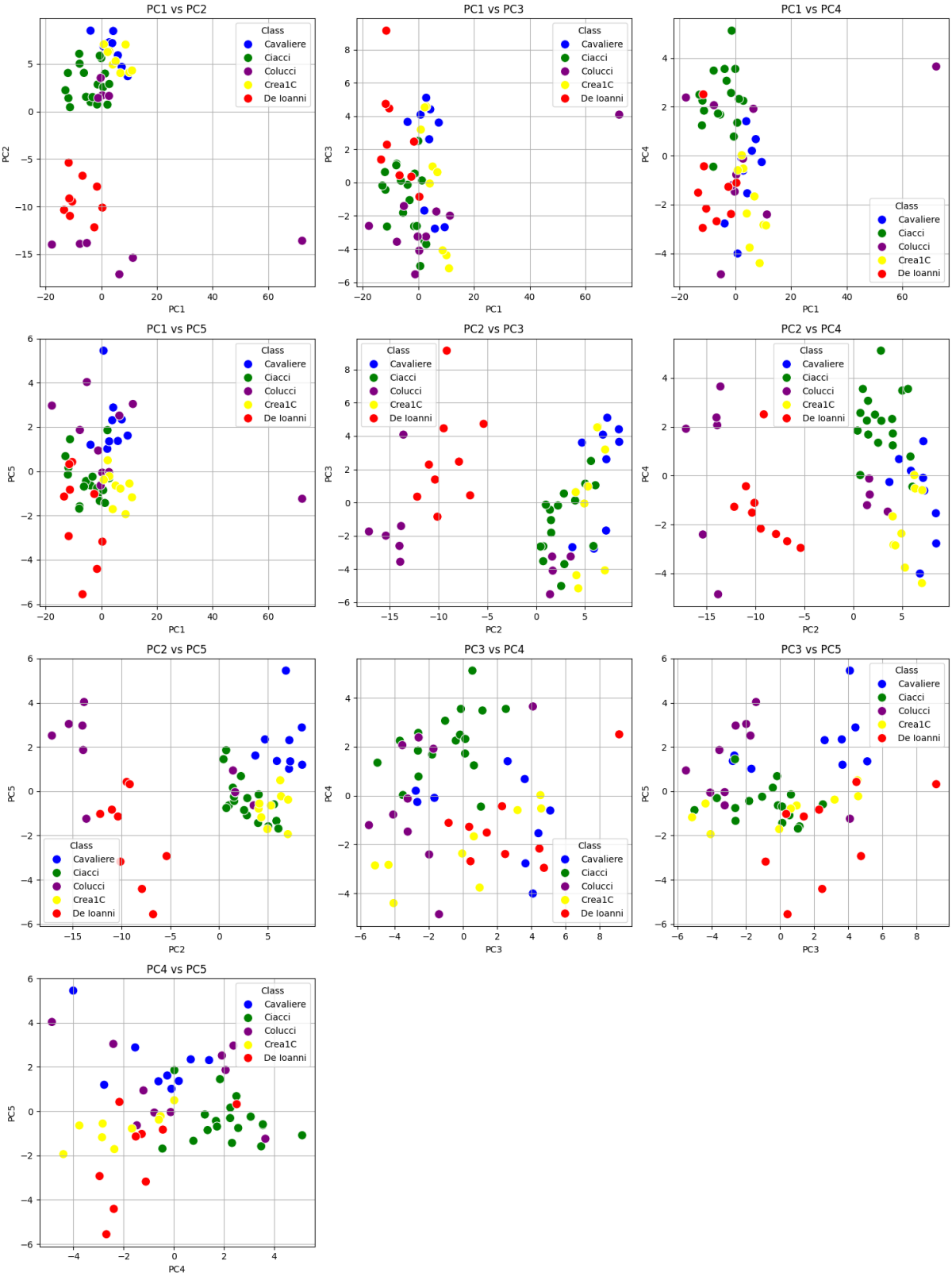


Figura 4.16 - Triticum durum (varietà Senatore Cappelli) - Distribuzione delle osservazioni nello spazio bidimensionale considerando la combinazione delle prime 5 PC.

Nel piano PC2 vs PC5, i gruppi Colucci e De Ioanni continuano a mantenere una distinzione visiva, mentre si osserva un incremento della dispersione per Cavaliere e Ciacci. In particolare, Crea1C appare più variabile e diffuso nello spazio.
Le proiezioni PC3 vs PC4 e PC3 vs PC5 evidenziano una maggiore sovrapposizione tra i gruppi, riducendo la capacità discriminativa di queste componenti. Tuttavia, De Ioanni mantiene una certa coerenza interna con minore dispersione.
Infine, il piano PC4 vs PC5 risulta il meno informativo ai fini della discriminazione, mostrando un'ampia sovrapposizione tra tutti i gruppi aziendali, confermando che le sole componenti PC4 e PC5 non sono sufficientemente descrittive per separare i campioni in base all'origine.

Addestramento dei modelli di machine learning supervisionato

I risultati riportati nella Tabella 4.8 illustrano le performance predittive ottenute durante la fase di addestramento dei modelli di classificazione. L'analisi è stata condotta mediante validazione incrociata stratificata a 3 fold, in presenza di rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza dei modelli rispetto a una variabilità simulata. Per ciascun algoritmo sono riportate l'accuratezza media, la deviazione standard e gli iperparametri ottimali individuati tramite ottimizzazione con GridSearchCV.

Tabella 4.8 - Triticum durum (varietà Senatore Cappelli) - Prestazioni predittive dei modelli di classificazione durante l'addestramento con dati affetti da rumore gaussiano.

Modello	Accuracy (mean)	Accuracy (std)	Parametri ottimali
Random Forest	0.909	0.025	{'max_depth': None, 'n_estimators': 200}
SVM	0.909	0.025	{'C': 0.1, 'kernel': 'linear'}
XGBoost	0.816	0.071	{'learning_rate': 0.1, 'max_depth': 3, 'n_esti...
k-NN	0.799	0.030	{'metric': 'manhattan', 'n_neighbors': 3}
Naive Bayes	0.855	0.069	default
Decision Tree	0.765	0.061	{'max_depth': None}
Logistic Regression	0.908	0.0276	{'C': 1}

Nel complesso, le performance risultano elevate e relativamente stabili, con valori di accuratezza media compresi tra 0.765 (Decision Tree) e 0.909 (Random Forest e SVM). Logistic Regression ha mostrato risultati comparabili (0.908), mentre modelli come XGBoost (0.816), Naive Bayes (0.855) e k-NN (0.799) hanno riportato performance leggermente inferiori, ma comunque soddisfacenti. La variabilità osservata può essere attribuita sia alle caratteristiche intrinseche degli algoritmi, sia alla perturbazione artificiale introdotta nei dati tramite l'aggiunta di rumore, in una fase intermedia del workflow analitico.

Le capacità di generalizzazione dei modelli saranno successivamente valutate attraverso la classificazione di campioni esterni non affetti da rumore, al fine di stimarne l'effettiva efficacia in condizioni operative realistiche.

Nei grafici seguenti sono riportate le metriche di performance dei modelli ottenute durante la fase di validazione su dati non precedentemente visti, affetti da rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza predittiva degli algoritmi in condizioni di lieve perturbazione.

La Figura 4.17 presenta un confronto diretto tra i modelli in termini di accuratezza media, evidenziando la percentuale complessiva di classificazioni corrette rispetto al totale delle osservazioni, mentre la Figura 4.18 riporta i valori di F1-score macro, metrica particolarmente utile per valutare la capacità del modello di bilanciare precisione e recall nelle classificazioni multi-classe. La Figura 4.19, infine, rappresenta la relazione tra precisione e recall macro, e permette di valutare l'equilibrio tra la capacità del modello di non produrre falsi positivi (precisione) e quella di rilevare correttamente le classi positive (recall).

L'analisi congiunta di queste metriche evidenzia una sostanziale coerenza con quanto osservato in fase di addestramento, confermando l'elevata capacità discriminativa della maggior parte dei modelli anche in presenza di una lieve perturbazione dei dati. In particolare, algoritmi come SVM, Random Forest e Logistic Regression mantengono livelli di accuratezza e F1-score macro molto elevati (superiori a 0.90), confermando la loro solidità. Tali modelli mostrano inoltre un ottimo equilibrio tra precisione e recall, posizionandosi nella zona superiore destra dello spazio metrico, indicativa di prestazioni bilanciate e robuste.

Altri algoritmi, pur riportando valori leggermente inferiori, mostrano comunque performance soddisfacenti e stabili, compatibili con una buona capacità generalizzante e una discreta affidabilità in presenza di rumore contenuto.

Nel complesso, i risultati ottenuti suggeriscono che la classificazione dei campioni in esame possa essere agevolata da una buona separabilità dei dati metabolomici, e che i modelli sviluppati presentano una robustezza predittiva elevata, potenzialmente applicabile in contesti operativi reali. In ogni caso, la successiva validazione su set esterni non affetti da rumore costituirà un passaggio fondamentale per la valutazione definitiva della generalizzabilità dei modelli.

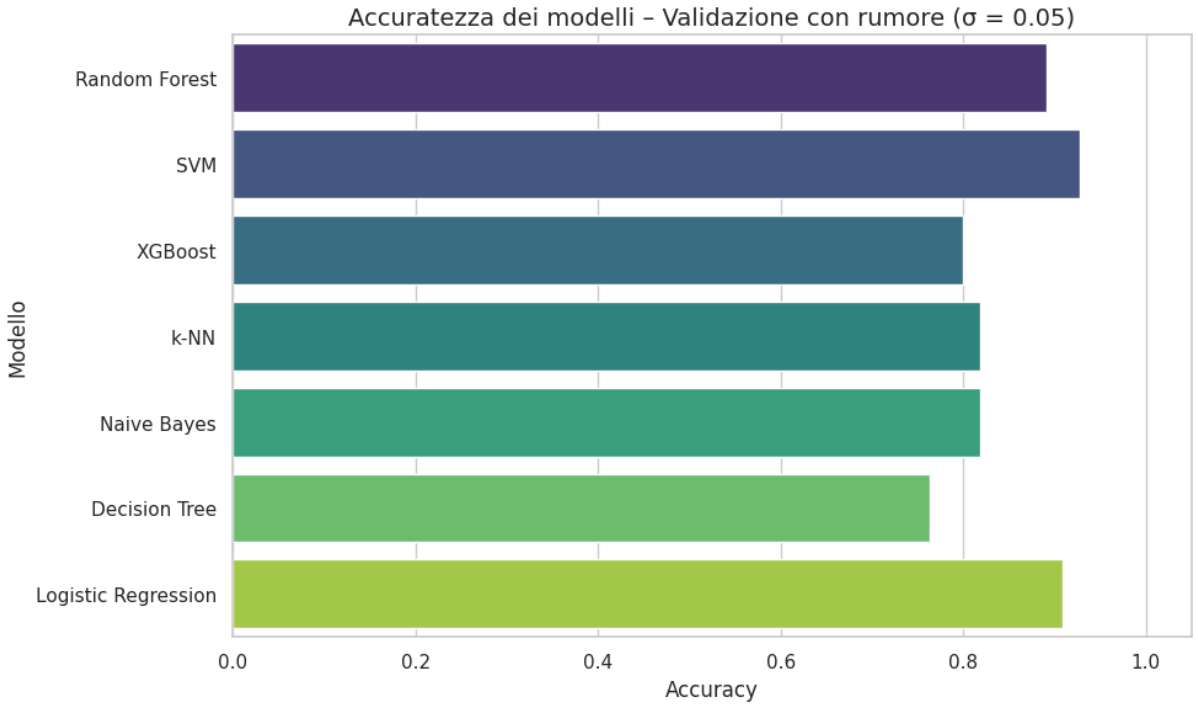


Figura 4.17 - Triticum durum (varietà Senatore Cappelli) - Accuracy per ciascun modello durante predizione su dati non visti.

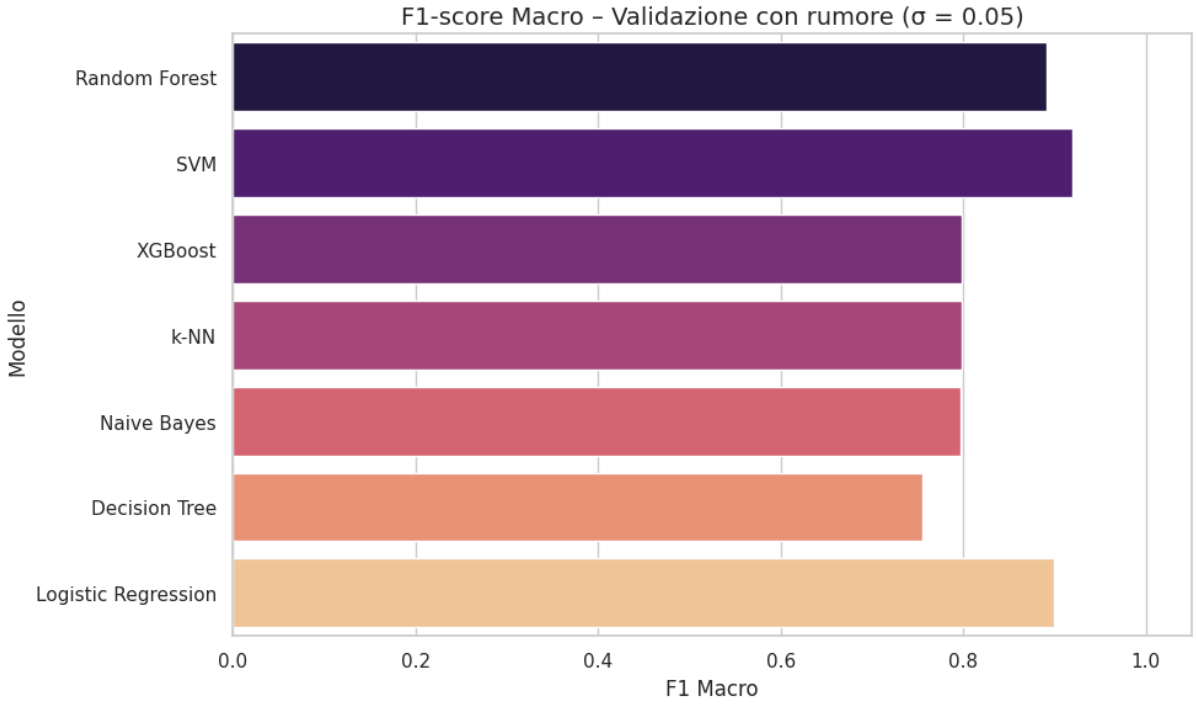


Figura 4.18 - Triticum durum (varietà Senatore Cappelli) - F1-score Macro per ciascun modello durante predizione su dati non visti.

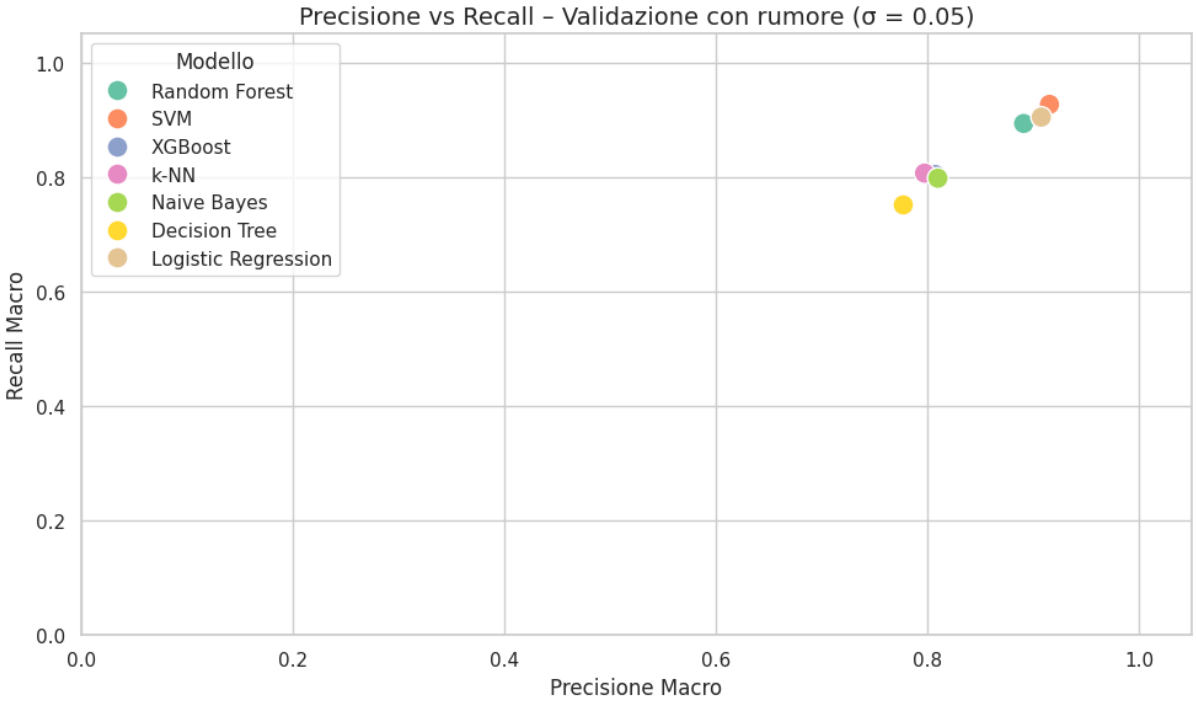


Figura 4.19 - Triticum durum (varietà Senatore Cappelli) - Precisione vs Recall macro per ciascun modello durante predizione su dati non visti.

Classificazione finale e selezione delle feature più rilevanti

I risultati riportati nella Tabella 4.9 evidenziano le prime 20 variabili (buckets NMR) che hanno mostrato la maggiore rilevanza durante il processo di classificazione condotto dai diversi algoritmi di machine learning, valutata mediante il criterio della *Mean Decrease Accuracy* (MDA). L'analisi si riferisce alla fase di test dei modelli eseguita su dati affetti da rumore gaussiano di intensità moderata ($\sigma = 0.1$), al fine di simulare condizioni realistiche di variabilità sperimentale.

Le variabili sono elencate in ordine decrescente rispetto al valore massimo di importanza MDA registrato, e sono accompagnate dagli intervalli spettrali corrispondenti (espressi in ppm) e dal numero di modelli predittivi nei quali sono risultate significative per la classificazione.

Tabella 4.9 – Triticum durum (varietà Senatore Cappelli) - Principali variabili (buckets NMR) selezionate tramite MDA durante il processo di classificazione.

Num	Variabile	PPM		Totale modelli	Importanza MDA
1	Var120	4.62	±0.02	6	1.398
2	Var121	4.58	±0.02	6	1.307
3	Var110	5.22	±0.02	7	1.175
4	Var203	1.30	±0.02	7	1.157
5	Var202	1.34	±0.02	7	1.141
6	Var107	5.34	±0.02	7	1.137
7	Var213	0.90	±0.02	7	1.041
8	Var184	2.06	±0.02	7	1.036
9	Var179	2.26	±0.02	7	1.031
10	Var166	2.78	±0.02	7	1.014
11	Var150	3.42	±0.02	7	0.986
12	Var201	1.38	±0.02	7	0.971
13	Var145	3.62	±0.02	3	0.933
14	Var108	5.30	±0.02	6	0.905
15	Var196	1.58	±0.02	6	0.887
16	Var159	3.06	±0.02	2	0.884
17	Var162	2.94	±0.02	2	0.882
18	Var146	3.58	±0.02	7	0.872
19	Var227	0.34	±0.02	6	0.863
20	Var156	3.18	±0.02	5	0.850

Le prime 20 variabili selezionate mostrano valori di MDA compresi tra 0.850% e 1.398%, a indicare una distribuzione dell'importanza relativamente ampia, con alcune variabili che esercitano un'influenza più marcata rispetto ad altre.

In particolare, Var120 (4.62 ppm) e Var121 (4.58 ppm) si distinguono per i valori MDA più elevati, pari rispettivamente a 1.398% e 1.307%, ed entrambe sono state selezionate da sei modelli su sette. La loro importanza risulta quindi solida, pur non emergendo in tutti gli algoritmi. A questi segnali si aggiungono Var110 (5.22 ppm), Var203 (1.30 ppm) e Var202 (1.34 ppm), tutte selezionate da tutti e sette i modelli, con MDA compresi tra 1.141% e 1.175%, a testimonianza di un ruolo centrale e ricorrente nei processi di classificazione.

È interessante notare che ben dieci variabili sono state individuate come rilevanti da tutti e sette i modelli. Tra queste, oltre alle già citate, figurano anche Var107 (5.34 ppm), Var213 (0.90 ppm), Var184 (2.06 ppm), Var179 (2.26 ppm), Var166 (2.78 ppm), Var150 (3.42 ppm), Var201 (1.38 ppm) e Var146 (3.58 ppm). Sebbene alcune di esse presentino valori di MDA inferiori a

1%, la loro costante selezione conferma una robustezza trasversale e una probabile rilevanza sistemica nel pattern metabolico discriminante.

Al contrario, altre variabili, come Var145 (3.62 ppm), Var159 (3.06 ppm) e Var162 (2.94 ppm), pur mostrando valori di MDA tra 0.882% e 0.933%, sono state selezionate solo da due o tre modelli, indicando una rilevanza limitata o condizionata dalla struttura dell'algoritmo. Tali segnali, seppur non trascurabili, dovrebbero essere interpretati con maggiore cautela, in quanto meno stabili su base inter-modello.

Validazione mediante set di dati esterno

Nella tabella seguente è riportata la performance dei diversi modelli in termini di accuratezza nella predizione su un set di dati esterno, privo di rumore gaussiano, ottenuto come descritto nel paragrafo 3.2.2.1.

In Figura 4.20 viene presentata, a titolo esemplificativo, la matrice di confusione relativa al modello Random Forest ottenuta nella fase di predizione sul set esterno.

Tabella 4.10 – Triticum durum (varietà Senatore Cappelli) - Accuratezza, F1-score macro e F1-score weighted dei modelli di classificazione applicati al set esterno di dati in ambito di lavoro Senatore Cappelli

Modello	Accuracy	F1_macro	F1_weighted
Random Forest	1,000	1,000	1,000
SVM	0,964	0,963	0,964
XGBoost	0,946	0,948	0,946
k-NN	0,964	0,963	0,964
Naive Bayes	0,909	0,907	0,906
Decision Tree	1,000	1,000	1,000
Logistic Regression	0,982	0,985	0,982

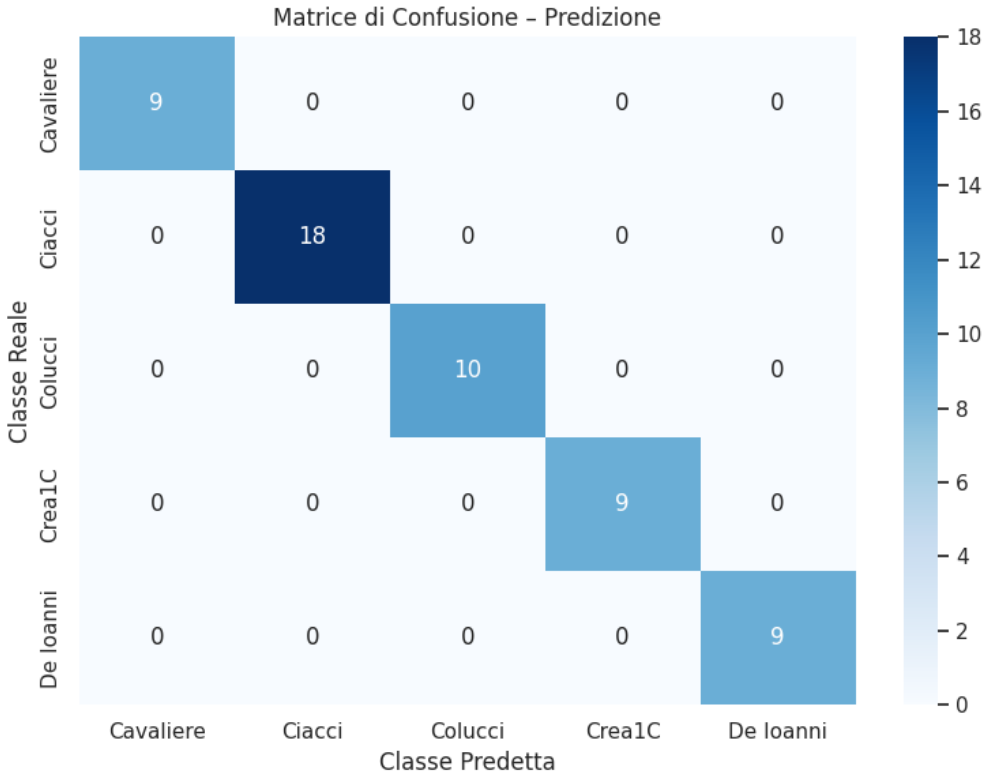


Figura 4.20 – Triticum durum (varietà Senatore Cappelli) - Matrice di confusione ottenuta con il modello Random Forest durante classificazione con set esterno di dati.

I risultati mostrano come tutti gli algoritmi raggiungano performance molto elevate, con accuratezze superiori al 90%, a conferma della buona qualità delle soluzioni sviluppate. In particolare, i modelli Random Forest e Decision Tree hanno ottenuto una classificazione perfetta (Accuracy = 1.000), senza alcun errore di predizione. Anche Logistic Regression, SVM e k-NN hanno mostrato prestazioni eccellenti, con accuratezze rispettivamente pari a 0.982, 0.964 e 0.964, e valori di F1-score elevati e bilanciati, indicativi di un'ottima capacità discriminativa. Il modello XGBoost si è attestato su un'accuratezza di 0.946, mantenendo un buon equilibrio tra precisione e richiamo, mentre Naive Bayes ha mostrato prestazioni leggermente inferiori (Accuracy = 0.909), pur restando sopra la soglia del 90%.

La matrice di confusione ottenuta dal modello Random Forest fornisce un'assenza totale di errori di classificazione: tutti i campioni sono stati correttamente assegnati alla classe di appartenenza. Questo risultato dimostra che l'addestramento su dati perturbati ha favorito la costruzione di modelli robusti, generalizzabili e adattabili a dati non manipolati.

Discussione finale

L'approccio automatizzato basato sull'integrazione tra PCA e modelli di machine learning supervisionato ha mostrato un'elevata efficacia nella discriminazione tra campioni di *Triticum durum* (var. Senatore Cappelli) provenienti da diverse aziende agricole.

I modelli predittivi addestrati sulle componenti latenti, in presenza di rumore gaussiano ($\sigma = 0.1$), hanno evidenziato performance elevate. In particolare, modelli come Logistic Regression, Random Forest e SVM hanno ottenuto una accuratezza > 0.90, mantenendo buoni livelli di F1-score e stabilità anche in presenza di perturbazione.

La selezione delle feature tramite MDA ha individuato diverse variabili condivise da tutti i modelli, confermando la presenza di segnali discriminanti robusti e coerenti tra algoritmi.

I risultati ottenuti su dati esterni, privi di rumore, hanno ulteriormente confermato l'efficacia del protocollo: Random Forest e Decision Tree hanno raggiunto una classificazione perfetta, mentre gli altri modelli hanno mantenuto accuratezze superiori al 90%, dimostrando un'elevata capacità di generalizzazione. La totale assenza di errori nella matrice di confusione di Random Forest evidenzia la solidità del modello e la validità dell'approccio adottato.

4.2.4 Confronto tra approccio classico e automatizzato

Il confronto tra l'approccio chemiometrico convenzionale e quello automatizzato AI-driven applicato ai campioni di Triticum durum varietà "Senatore Cappelli" ha messo in luce differenze metodologiche rilevanti, ma anche una sostanziale convergenza nei risultati, con vantaggi complementari in termini di interpretabilità e capacità predittiva.

Dal punto di vista esplorativo, l'analisi PCA condotta con metodo classico ha permesso di evidenziare una chiara separazione tra i campioni in funzione dell'origine geografica. In particolare, le componenti PC2 e PC3 hanno mostrato una forte capacità discriminante, separando nettamente i gruppi aziendali De Ioanni, Colucci e Ciacci. L'analisi dei contributi percentuali ha permesso di selezionare le variabili più influenti, successivamente confermate tramite ANOVA e test post-hoc di Tukey. L'assegnazione putativa dei bucket a metaboliti noti – tra cui glucosio, saccarosio, fruttosio, serina, etanolamina e alanina – ha fornito una lettura biochimica coerente della variabilità osservata, sottolineando l'importanza dei metaboliti zuccherini e azotati nella differenziazione geografica dei campioni. L'approccio automatizzato, fondato sull'integrazione tra PCA e modelli di machine learning supervisionato, ha mostrato performance eccellenti nella fase predittiva. I modelli Random Forest, SVM, Logistic Regression e Decision Tree hanno raggiunto un'accuratezza pari o superiore al 96% nella classificazione su dati esterni, con F1-score macro e weighted estremamente elevati. In

particolare, Random Forest e Decision Tree hanno ottenuto una classificazione perfetta (accuracy = 1.000), senza alcun errore di assegnazione. La selezione delle feature mediante MDA ha individuato un set di variabili altamente informative, con dieci segnali selezionati da tutti i modelli e valori di importanza compresi tra 0.85% e 1.40%. Tra queste, Var110 (5.22 ppm), Var203 (1.30 ppm) e Var150 (3.42 ppm) si sono rivelate centrali nel processo di classificazione. Un aspetto particolarmente significativo emerso dal confronto riguarda la coerenza tra le variabili selezionate nei due approcci. L'analisi classica e quella automatizzata convergono su diversi segnali chiave, in particolare nella regione spettrale δ 1.30–5.40 ppm. Ad esempio, Var110, associata al glucosio (H1), è risultata significativa sia nel test di Tukey sia nel ranking MDA, mentre Var203 e Var202, localizzate nell'area degli amminoacidi alifatici, sono state selezionate da tutti i modelli e hanno mostrato importanza statistica anche nell'analisi PCA tradizionale. Tale sovrapposizione indica una convergenza tra metodi diversi ma complementari, che rafforza la validità delle variabili identificate. Inoltre, l'approccio AI-driven ha evidenziato una maggiore capacità di gestione della complessità multivariata, grazie all'introduzione del rumore controllato e all'utilizzo di tecniche di cross-validation. Questi accorgimenti hanno permesso di sviluppare modelli robusti e generalizzabili, adatti a scenari applicativi reali. Al contrario, il metodo classico, pur offrendo una lettura più immediata e biologicamente interpretabile, risulta meno adatto alla costruzione di strumenti predittivi autonomi. I risultati ottenuti suggeriscono che l'approccio automatizzato possa rappresentare un'evoluzione del metodo classico, mantenendo la coerenza con le interpretazioni biochimiche tradizionali, ma potenziandole con capacità predittive, scalabilità e automatizzazione del processo analitico.

4.2.5 Esempio di serializzazione e generazione hash per registrazione del fingerprint metabolico su blockchain

Di seguito viene mostrato un esempio di file serializzato in formato JSON ottenuto a partire dal segnale NMR grezzo (FID) acquisito da un campione rappresentativo della matrice Triticum durum – varietà Senator Cappelli (codice 12_2NMR_CIACCI_1_I_r3_D2O), attraverso spettroscopia 1H HR-MAS.


```
{
  "sample_id": "Ciacci_1_I_r3",
  "matrix": "Senatore Cappelli",
  "instrument": "av400",
  "operator": "Cangemi",
  "date": "2023-05-26T11:17:39",
  "temperature": 297.9867,
  "pulse_program": "zgpr",
  "fid_real": [
    0.0,
    0.0,
…
    -492.0
  ],
  "fid_imag": [
    0.0,
    0.0,
…
    -118.0
  ],
  "is_domain": "time",
  "timestamp": "2025-06-10T06:25:00.032402"
}
```


Figura 4.21 – Triticum durum (varietà Senatore Cappelli) - Esempio di serializzazione da FID Bruker

Il file JSON include la parte reale e quella immaginaria del FID e un insieme di metadati contestuali: strumento, operatore, data, temperatura, sequenza di impulso e timestamp.

Il contenuto del file JSON, sottoposto a funzione di hash SHA256, ha fornito il seguente identificativo digitale:

bfd5834463b907f8d72f2677ab0fb5d0e7ef3bd0afad8f20980d8a572638fcf6

Figura 4.22 - Triticum durum (varietà Senatore Cappelli) - Esempio di generazione hash da FID Bruker

L’hash così ottenuto può fungere da identificativo digitale non modificabile, utile per ancorare il dato analitico a uno smart contract distribuito.
Questo hash rappresenta una "firma digitale" univoca del file, che ne garantisce l’integrità e l’autenticità. Una volta registrato su una rete blockchain (es. Hyperledger Fabric), tale impronta permette di verificare, in qualsiasi momento, che il file non sia stato modificato e che l’acquisizione sia avvenuta in condizioni sperimentali note e tracciabili.

4.3 Concentrato di Pomodoro

4.3.1 Descrizione dei profili spettrali

L'analisi mediante spettroscopia 1H HR-MAS NMR ha consentito di acquisire spettri di alta qualità per i campioni di concentrato di pomodoro provenienti da Italia, Cina e Spagna, permettendo di indagare in modo dettagliato la composizione metabolica di ciascuna matrice. Gli spettri ottenuti, riportati in Figura 4.23, evidenziano una sostanziale sovrapposizione qualitativa tra i diversi campioni, con una distribuzione dei segnali complessivamente simile lungo l'intervallo chimico compreso tra 0 e 10 ppm. Tale uniformità riflette la natura relativamente omogenea del prodotto alimentare considerato, il quale, pur derivando da filiere produttive differenti, mantiene una base metabolica comune dovuta alla medesima matrice vegetale di origine. Nonostante l'assenza di marcate differenze visive a livello globale, un'osservazione delle principali regioni spettrali consente di evidenziare segnali riconducibili ad acidi organici e amminoacidi nel range compreso tra 0.5 e 2.5 ppm (Savorani et al., 2009). La zona centrale dello spettro, compresa tra 2.5 e 6.0 ppm, risulta dominata da un fitto pattern di segnali sovrapposti, indicativi della presenza abbondante di carboidrati solubili, in particolare fruttosio, glucosio e saccarosio, la cui intensità e distribuzione variano lievemente in funzione della provenienza geografica e delle modalità di trasformazione industriale. Infine, nella regione compresa tra 6.0 e 9.0 ppm si collocano i segnali attribuibili ai composti fenolici e aromatici. In tale intervallo, un'osservazione ravvicinata – facilitata dall'ingrandimento riportato nel riquadro a sinistra della figura – suggerisce la presenza di lievi differenze in termini di intensità tra i campioni analizzati. Tuttavia, pur in assenza di un'evidente separazione tra i profili spettrali osservabili ad occhio nudo, l'organizzazione dei segnali nelle diverse aree chimiche e le sottili variazioni nella loro intensità relativa forniscono indicazioni preliminari sulla possibile esistenza di tratti distintivi associabili alla regione di provenienza.

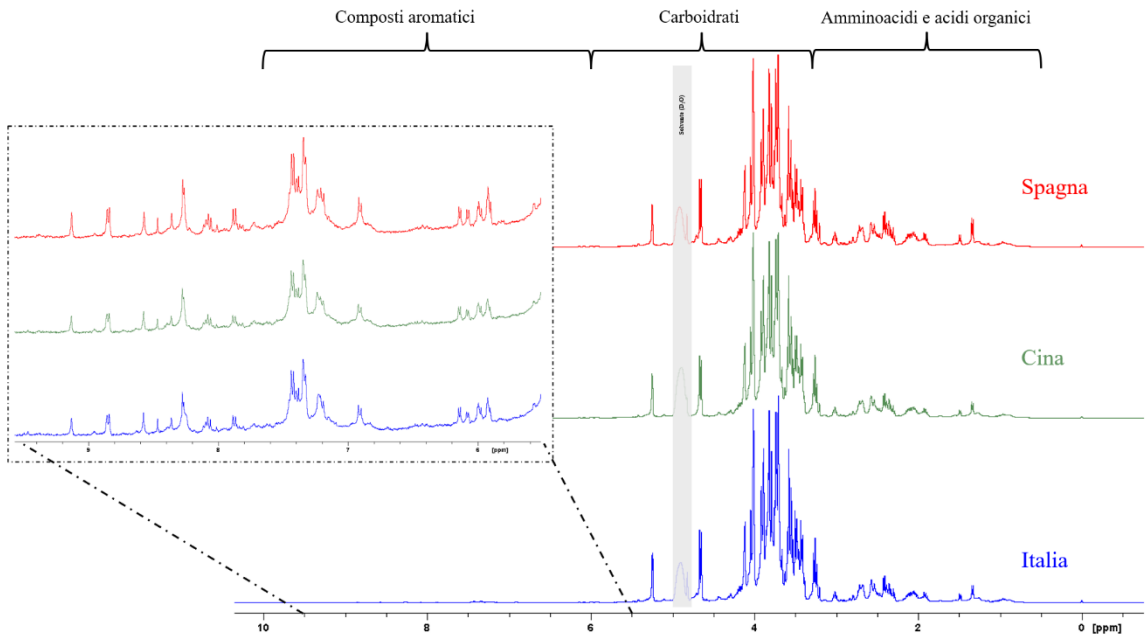


Figura 4.23 – Concentrato di Pomodoro - Spettri 1H-NMR.

4.3.2 Risultati dell'analisi chemiometrica convenzionale

PCA

La Figura 4.24 mostra due esempi rappresentativi di score plot ottenuti dall'analisi PCA eseguita sui campioni di concentrato di pomodoro mediante il software XLSTAT. Nella proiezione bidimensionale definita dalle prime due componenti principali (PC1 e PC2) è possibile osservare una parziale separazione tra i gruppi campionari, con i campioni di origine spagnola che tendono a distribuirsi lungo valori positivi della prima componente, mentre quelli italiani e cinesi risultano collocati principalmente nella porzione sinistra del piano. In particolare, i concentrati provenienti dalla Cina mostrano una maggiore dispersione lungo l'asse PC2, suggerendo una maggiore eterogeneità metabolica interna rispetto agli altri gruppi. I campioni italiani si distribuiscono in modo più compatto, collocandosi prevalentemente nella regione intermedia, a cavallo tra i segnali caratteristici delle altre due provenienze.

La rappresentazione tridimensionale nel piano definito da PC1, PC2 e PC3 consente di esplorare ulteriormente la distribuzione spaziale dei campioni, evidenziando la presenza di una maggiore varianza intraclasse nei campioni cinesi e una certa tendenza alla separazione tra i gruppi geografici. In questo spazio multivariato, i campioni spagnoli appaiono più raccolti e orientati lungo la direzione di PC1, mentre quelli italiani si dispongono trasversalmente, mostrando una

certa sovrapposizione con entrambe le altre provenienze. La configurazione osservata suggerisce la possibile esistenza di segnali metabolici caratteristici associati all'area di produzione, ma anche una certa sovrapposizione tra i profili, compatibile con le osservazioni già discusse nella sezione spettrale.

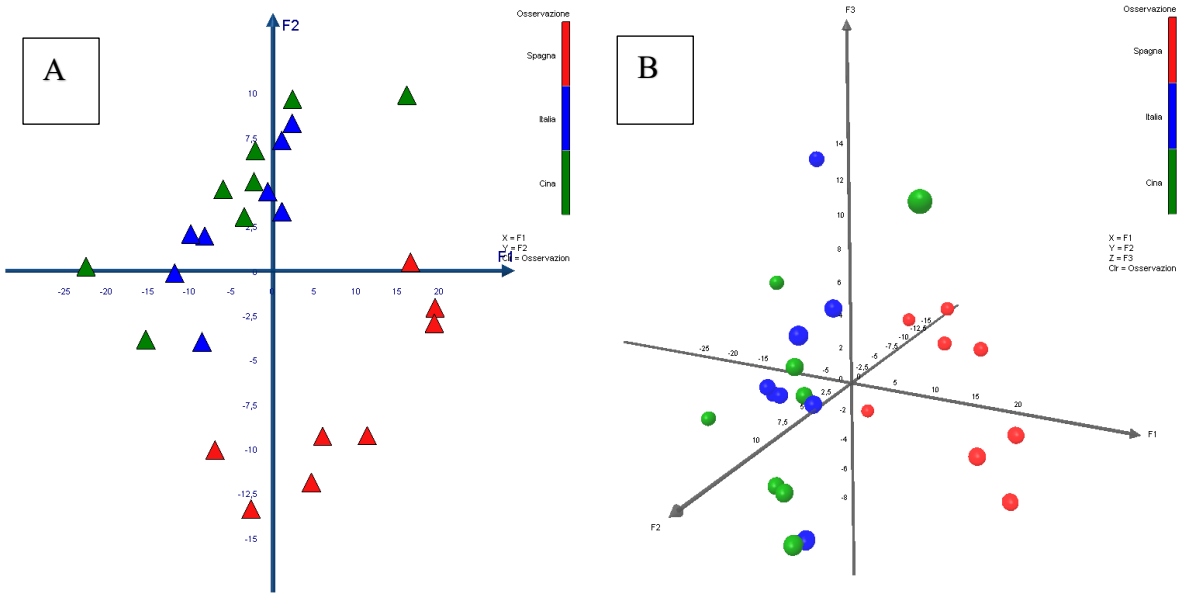


Figura 4.24– Concentrato di Pomodoro - Score Plot da PCA.

Selezione delle variabili significative

La Tabella 4.11 riporta, per ciascuna delle prime sette componenti principali, l'elenco delle variabili che risultano significativamente differenti tra almeno due gruppi ($p < 0.05$), sulla base dell'analisi della varianza univariata (ANOVA). È opportuno precisare che la significatività statistica indicata non implica che ciascuna variabile discrimini tutte le classi tra loro, bensì che, per almeno un confronto tra coppie di gruppi, è stata rilevata una differenza significativa.

Per una maggiore chiarezza interpretativa, è stata successivamente condotta un'analisi post-hoc (test di Tukey HSD) al fine di individuare con precisione quali gruppi differiscono tra loro per ciascuna variabile significativa. Nella presente tabella è tuttavia riportato esclusivamente l'elenco complessivo delle variabili che hanno mostrato una significatività globale (ANOVA a $p < 0.05$), senza entrare nel dettaglio dei confronti specifici tra classi. Tale approccio è stato adottato con l'obiettivo di fornire una sintesi preliminare delle variabili maggiormente coinvolte nella separazione dei gruppi lungo ciascuna componente principale.

Tabella 4.11 – Concentrato di Pomodoro - Elenco delle variabili significative.

	Selezione PCA+ANOVA+Tukey
PC	Variabili significative ($p < 0.05$)
PC1	Var29, Var39, Var40, Var41, Var42, Var43, Var44, Var45, Var51, Var52, Var53, Var67, Var69, Var92, Var93, Var190, Var191, Var192, Var193, Var195, Var197, Var205, Var206, Var207, Var208, Var209, Var216
PC2	Var4, Var107, Var108, Var109, Var148, Var149, Var150, Var154
PC3	Var135, Var138, Var139, Var140, Var141, Var142, Var143, Var146, Var147, Var148, Var149, Var150, Var154
PC4	Var169
PC5	Var110, Var159, Var169
PC6	Var107, Var110, Var156, Var158, Var196
PC7	Var169, Var196, Var216

Assegnazione delle variabili discriminanti ai metaboliti

Nella Tabella 4.12 è riportata l'assegnazione delle principali variabili (bucket spettrali) ai metaboliti putativi, effettuata mediante confronto con la libreria metabolomica precedentemente predisposta.

L'assegnazione spettrale, effettuata tramite confronto con database on-line e dati HSQC evidenzia quanto segue. Un primo gruppo di segnali, localizzati nella regione compresa tra circa 7.5 e 8.5 ppm, comprende le variabili Var29, Var39–Var45 e Var51–Var53, tutte appartenenti alla prima componente principale (PC1). Questi bucket sono stati assegnati a protoni aromatici riconducibili a composti azotati come la trigonellina, l'adenina, l'uridina, e ad amminoacidi aromatici quali triptofano, tirosina e fenilalanina. La presenza della trigonellina, in particolare, è stata confermata da più bucket (Var4, Var29, Var40, Var42) e rappresenta un importante indicatore della maturazione del frutto, nonché un marcatore metabolico frequentemente rilevato nei prodotti vegetali. Analogamente, la rilevazione di amminoacidi aromatici come triptofano e tirosina può suggerire un coinvolgimento di vie biosintetiche legate alla risposta a stress o al grado di trasformazione termica del prodotto.

Tabella 4.12 – Concentrato di Pomodoro - Assegnazione putativa delle variabili ai metaboliti.

Buckets	PC	δ ^{1}H (ppm)	δ ^{13}C (ppm)	Metabolita Assegnato
Var197	PC1	1,54		CH_2–CH_2–COOH (acidi)
Var196	PC6/PC7	1,58		GABA, Acidi grassi
Var195	PC1	1,62		Ile / GABA / Gln
Var193	PC1	1,7		Arg / Leu / Ile
Var192	PC1	1,74		Arg / Glu / Leu
Var191	PC1	1,78		GABA / Glu / acidi
Var190	PC1	1,82		GABA / ω-CH_2
Var169	PC4/PC5/PC7	2,66		CH_2-COO (malato, GABA, acidi grassi)
Var159	PC5	3,06		GABA (γ-CH_2), lattato
Var158	PC6	3,1		GABA / α-Fru / lattato
Var156	PC6	3,18		Acido glutammico (γ-H)
Var154	PC2/PC2	3,26		β-Glucosio / α-Glucosio
Var150	PC2	3,42		β-Glucosio (C4-H)
Var150	PC3	3,42		β-Glucosio (C4-H)
Var149	PC2/PC3	3,46		β-Glucosio (C3-H / C5-H)
Var148	PC3	3,5	78,58	α-/β-Glucosio, α-Glu, Fru
Var148	PC2	3,5	78,58	α-/β-Glucosio, α-Glu, Fru
Var147	PC3	3,54		β-Glu / α-Fru
Var146	PC3	3,58	73,96	Lattato, Glu, α-Fru
Var143	PC3	3,7		α-/β-Glucosio, lattato
Var142	PC3	3,74	75,41	α-Glu / α-Fru / Fru-F
Var141	PC3	3,78	57,06	Fru, Gln, β-Glu
Var140	PC3	3,82	70,35	α-Fru, Phe?
Var139	PC3	3,86	73,96	α-Fru / Glu
Var138	PC3	3,9	63,27	α-Fru, Gln, lattato
Var135	PC3	4,02	66,02	Lattato, α-Fru, GABA
Var110	PC5/PC6	5,22		Glicerolo (CH–O)
Var109	PC2	5,26	94,47	α-Glucosio (C1-H)
Var108	PC2	5,3		Amido / pectine
Var107	PC2/PC6	5,34		Acido α-galatturonico
Var69	PC1	6,86		Tirosina (H-3, H-5)
Var67	PC1	6,94		Tirosina (H-2, H-6)
Var53	PC1	7,5		Fenilalanina / Trp
Var52	PC1	7,54		Fenilalanina / Trp
Var51	PC1	7,58		Fenilalanina, Trp, Tyr
Var45	PC1	7,82		Trp, Tyr, derivati nucleosidici
Var44	PC1	7,86		Trp, derivati purinici
Var43	PC1	7,9		Trp, Tyr
Var42	PC1	7,94		Trigonellina, Trp
Var41	PC1	7,98		Trp, uridina
Var40	PC1	8,02		Trigonellina / adenosina
Var39	PC1	8,06		Adenina / Trigonellina
Var29	PC1	8,46		Trigonellina
Var4	PC2	9,46		Trigonellina

Un secondo insieme di variabili significative ricade nella regione degli zuccheri, compresa tra 3.2 e 5.4 ppm, principalmente associata alle componenti PC2, PC3 e PC6. Le variabili Var107–Var110, Var143–Var154 sono state assegnate a glucosio (α e β), fruttosio, acido galatturonico e residui di pectine e amido. In particolare, i bucket Var109 e Var110 mostrano valori chimici compatibili con protoni anomerici, mentre quelli compresi tra 3.4 e 3.9 ppm riflettono protoni ciclici di zuccheri riducenti. Questi composti sono noti per essere altamente rappresentati nel pomodoro e nei suoi derivati concentrati, poiché contribuiscono in modo rilevante al profilo sensoriale del prodotto (dolcezza, viscosità) e rappresentano un indicatore della varietà e del processo di lavorazione. La regione alifatica tra 1.5 e 2.0 ppm mostra un'elevata concentrazione di bucket significativi, tra cui Var190–Var197, tutti legati alla componente PC1. Questi segnali sono attribuiti a gruppi CH_2 di composti come GABA (acido γ-aminobutirrico), glutammato, acidi organici e amminoacidi a catena ramificata (isoleucina, leucina, valina). Il GABA, in particolare, è un importante metabolita con ruolo sia strutturale che funzionale nei vegetali, noto per accumularsi in risposta a condizioni di stress e frequentemente associato alla qualità nutrizionale del pomodoro maturo. Bucket come Var169 e Var196, presenti in più componenti principali (PC4–PC5–PC6–PC7), rafforzano l'importanza di questa regione spettrale come zona chiave per la discriminazione tra campioni. Infine, la presenza di segnali ancor più a monte dello spettro, come Var4 a 9.46 ppm, conferma ulteriormente la rilevanza della trigonellina come biomarcatore centrale in questo studio. Si tratta di un composto che, oltre ad avere una chiara identificazione spettrale, mostra coerenza con quanto riportato in letteratura riguardo la composizione dei derivati del pomodoro e il loro comportamento in risposta alla lavorazione industriale.

4.3.3 Risultati dell'analisi chemiometrica AI-driven

Importazione, verifica e pre-processing dei dati

La Figura 4.25 mostra un estratto della matrice dei bucket spettrali importata nello script, rappresentando i dati prima e dopo la standardizzazione mediante trasformazione Z-score.
Tale standardizzazione realizza una distribuzione centrata attorno allo zero e priva di dominanze numeriche, condizione ottimale per l'applicazione di modelli di machine learning supervisionato. Questo passaggio garantisce che ogni variabile contribuisca in egual misura al modello, prevenendo distorsioni dovute alla diversa scala delle intensità.

Dati Grezzi (prime 5 righe):

	replica	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	...	Var22
0	1	115026.12130	142361.81470	184526.79940	222692.4448	186179.46610	256749.2443	186074.83950	166133.37500	164629.07760	...	1554613.52
1	2	110882.54290	106552.78590	157126.65480	229365.7511	151973.96430	239437.9264	179979.37120	179780.35690	121915.49900	...	1130175.46
2	3	126513.44920	161353.26980	178617.09930	238359.5299	164444.69250	250415.1043	169966.89070	149743.40470	128995.27320	...	1464852.89
3	4	53292.29772	35783.10636	74381.76017	126945.0588	53703.94577	150426.8108	79838.83314	70407.29753	40827.99975	...	1127805.62
4	5	51186.43430	73204.27724	100957.64410	152075.9485	100272.98470	194070.8451	110963.14000	98606.32448	60360.77194	...	1496309.66

5 rows × 232 columns

Dati Standardizzati (prime 5 righe):

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224
0	0.898082	1.051364	0.865353	0.902375	1.204724	0.134489	0.452127	0.064140	1.208616	1.121497	...	0.157564	0.135190	0.263754	0.373715
1	0.774779	0.170052	0.258368	1.027895	0.499711	-0.154598	0.336935	0.319022	0.304138	0.197724	...	-1.821493	-1.899070	-1.866413	-1.933218
2	1.239917	1.518772	0.734438	1.197061	0.756747	0.028713	0.147719	-0.241972	0.454055	0.723954	...	-0.434851	-0.295017	-0.193168	-0.266705
3	-0.938968	-1.571694	-1.574648	-0.898559	-1.525742	-1.641019	-1.555518	-1.723718	-1.412926	-1.591512	...	-1.762600	-1.910428	-1.906636	-1.813175
4	-1.001633	-0.650704	-0.985922	-0.425866	-0.565903	-0.912195	-0.967332	-1.197050	-0.999311	-0.851381	...	-0.083419	-0.144250	-0.225545	-0.060952

5 rows × 230 columns

Figura 4.25 - Concentrato di Pomodoro –Estratto della matrice dei bucket spettrali prima e dopo la standardizzazione Z-score

Riduzione dimensionale mediante PCA

In analogia con quanto mostrato nel paragrafo 4.3.2, vengono presentati i risultati dell'analisi PCA condotta con l'impiego dello script. La Figura 4.26 riporta l'andamento della varianza cumulativa spiegata (Elbow Plot), evidenziando come la soglia del 95% venga raggiunta entro le prime cinque componenti principali. La combinazione dei criteri adottati — soglia del 95% della varianza spiegata, criterio di Kaiser e metodo del "Broken Stick" — ha condotto alla selezione delle prime cinque componenti principali, ritenute ottimali per l'addestramento e la predizione dei modelli di machine learning.

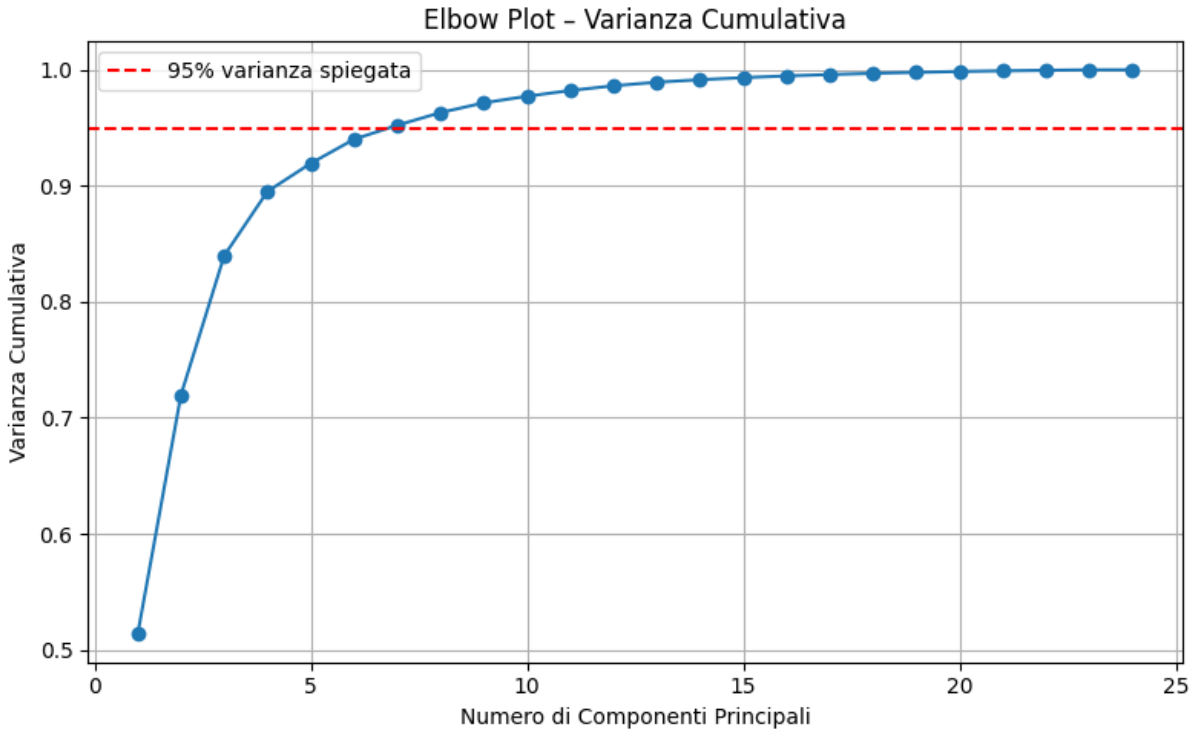


Figura 4.26 – Concentrato di Pomodoro - Elbow Plot relativo alla varianza cumulativa spiegata dalle componenti principali e indicazione della soglia al 95%.

Gli score plot bidimensionali ottenuti dall'analisi PCA automatizzata sono riportati in Figura 4.27.

L'analisi visiva delle combinazioni tra le componenti principali evidenzia che solo alcune componenti, in particolare PC1, PC2 e PC4, contribuiscono in modo significativo alla separazione tra le classi geografiche (Italia, Cina e Spagna). Le altre componenti, pur descrivendo parte della varianza, non mostrano un contributo discriminante evidente.

In particolare, nel piano PC1 vs PC2 si osserva una netta tendenza alla separazione della classe spagnola, i cui campioni si distribuiscono prevalentemente lungo valori positivi di PC1. I campioni italiani e cinesi si concentrano invece nella parte sinistra del grafico, risultando più sovrapposti. PC2 rafforza la separazione tra Italia e Spagna, evidenziando un gradiente di differenziazione anche lungo questo asse.

La componente PC4 contribuisce, seppur in maniera più sottile, a distinguere Italia e Cina, come osservabile nei piani PC2 vs PC4 e PC3 vs PC4, dove si intravede una leggera separazione tra queste due classi, non riscontrabile nei piani precedenti.

Le restanti combinazioni tra componenti principali (es. PC3 vs PC5, PC5 vs PC6, PC6 vs PC7) mostrano una sovrapposizione marcata tra le tre provenienze, suggerendo che queste

componenti rappresentano principalmente varianza intraclasse o rumore, piuttosto che informazione utile alla discriminazione tra gruppi.

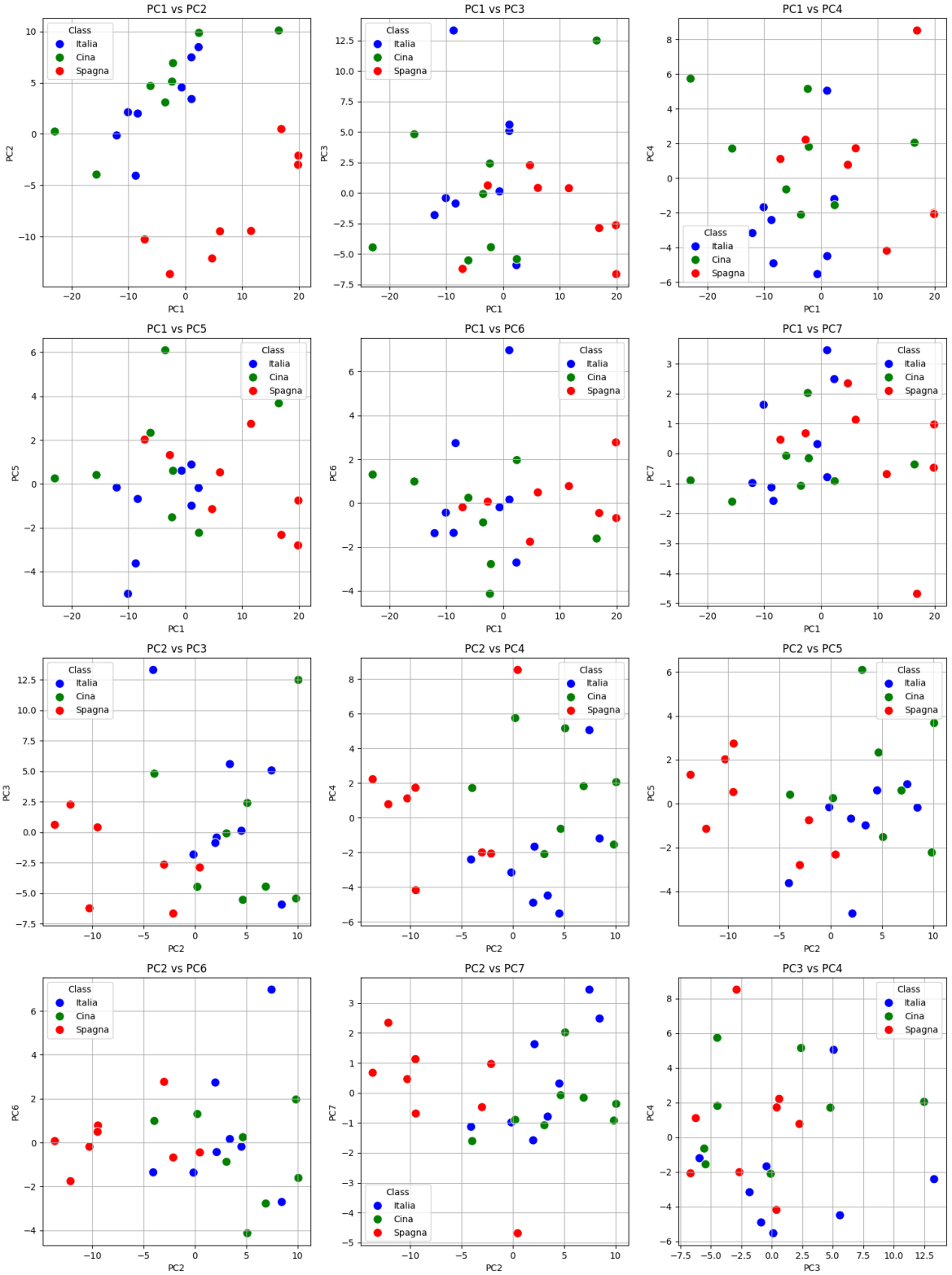


Figura 4.27 – Concentrato di Pomodoro - Distribuzione delle osservazioni nello spazio bidimensionale considerando la combinazione delle prime 7 PC (Parte 1 di 2)

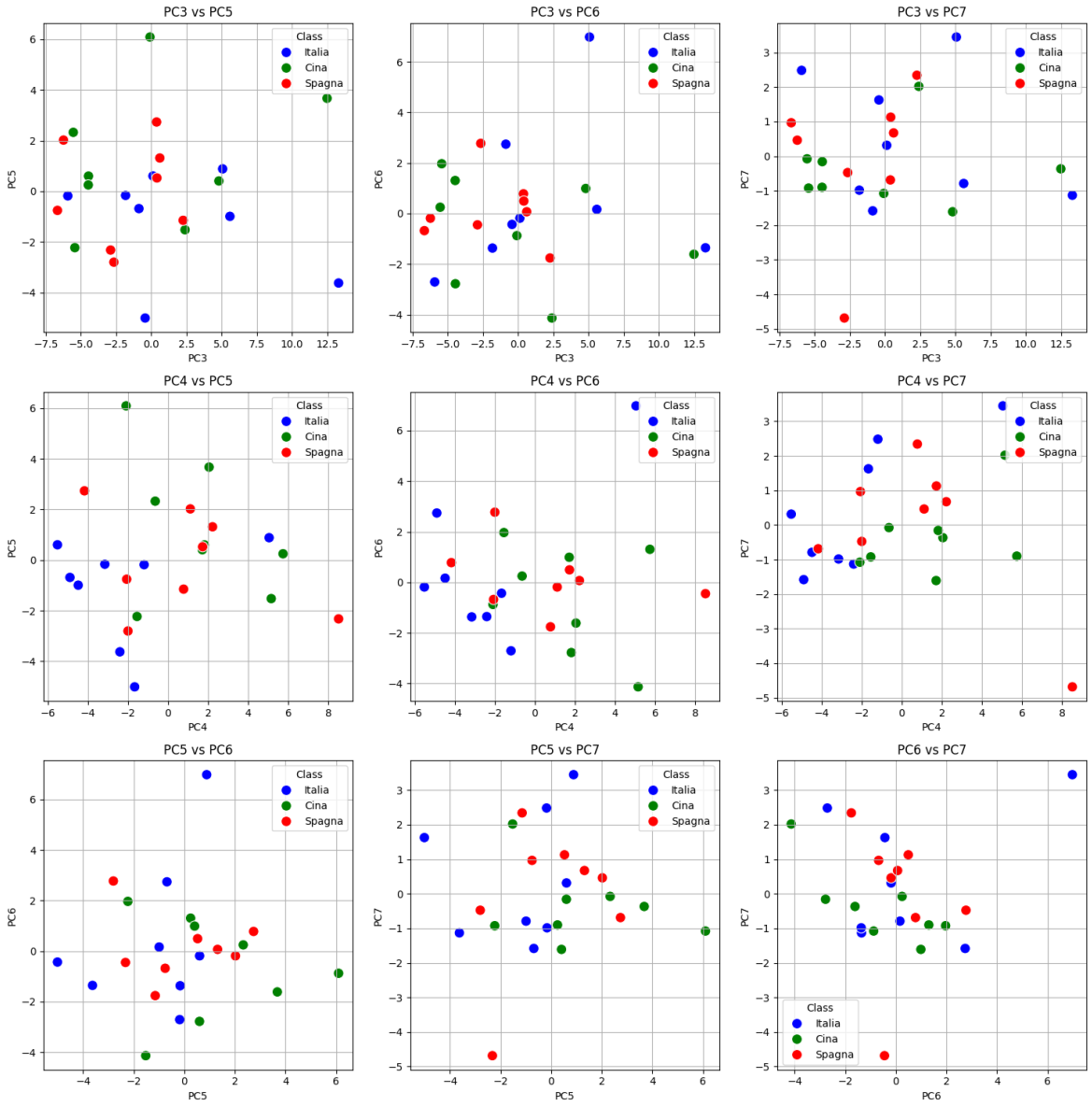


Figura 4.27 – Concentrato di Pomodoro - Distribuzione delle osservazioni nello spazio bidimensionale considerando la combinazione delle prime 7 PC (Parte 2 di 2)

Addestramento dei modelli di machine learning supervisionato

I risultati riportati nella Tabella 4.13 illustrano le performance predittive ottenute durante la fase di addestramento dei modelli di classificazione. L'analisi è stata condotta mediante validazione incrociata stratificata a 2 fold, in presenza di rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza dei modelli rispetto a una variabilità simulata. Per ciascun algoritmo sono riportate l'accuratezza media, la deviazione standard e gli iperparametri ottimali individuati tramite ottimizzazione con GridSearchCV.

Tabella 4.13 – Concentrato di Pomodoro - Prestazioni predittive dei modelli di classificazione durante l'addestramento con dati affetti da rumore gaussiano.

Modello	Accuracy (mean)	Accuracy (std)	Parametri ottimali
Random Forest	0.375	0.042	{'max_depth': None, 'n_estimators': 100}
SVM	0.792	0.042	{'C': 0.1, 'kernel': 'linear'}
XGBoost	0.375	0.042	{'learning_rate': 0.01, 'max_depth': 3.
k-NN	0.625	0.208	{'metric': 'euclidean', 'n_neighbors': 7}
Naive Bayes	0.625	0.125	default
Decision Tree	0.333	0.000	{'max_depth': None}
Logistic Regression	0.792	0.042	{'C': 1}

Nel complesso, le performance risultano piuttosto eterogenee, con valori di accuratezza compresi tra 0.333 e 0.792. Tale variabilità può essere attribuita sia alla complessità della matrice analizzata, sia alle caratteristiche intrinseche degli algoritmi, considerata anche la peculiarità dell'addestramento su set di dati perturbato artificialmente, in una fase intermedia del workflow. Le reali capacità di generalizzazione dei modelli verranno successivamente approfondite mediante la classificazione di campioni esterni non affetti da rumore.

Nei grafici seguenti sono riportate le metriche di performance dei modelli ottenute durante la fase di validazione su dati non precedentemente visti, affetti da rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza predittiva degli algoritmi in condizioni di lieve perturbazione.

La Figura 4.28 presenta un confronto diretto tra i modelli in termini di accuratezza media, evidenziando la percentuale complessiva di classificazioni corrette rispetto al totale delle osservazioni, mentre la Figura 4.29 riporta i valori di F1-score macro, metrica particolarmente utile per valutare la capacità del modello di bilanciare precisione e recall nelle classificazioni multi-classe. La Figura 4.30, infine, rappresenta la relazione tra precisione e recall macro, e permette di valutare l'equilibrio tra la capacità del modello di non produrre falsi positivi (precisione) e quella di rilevare correttamente le classi positive (recall).

L'analisi congiunta di queste metriche evidenzia una sostanziale coerenza con quanto osservato durante la fase di addestramento, confermando l'eterogeneità dei risultati tra i diversi algoritmi testati. Questa variabilità riflette le differenze strutturali tra i modelli e la loro diversa capacità di adattarsi a dati affetti da rumore, simulando scenari di lieve perturbazione. In particolare, SVM e Logistic Regression si confermano tra i modelli più robusti, mantenendo performance elevate e bilanciate anche in condizioni non ottimali. Al contrario, algoritmi come Random Forest e Decision Tree mostrano una maggiore sensibilità alla variabilità indotta, con una conseguente riduzione della capacità discriminativa. Tali risultati, sebbene informativi,

rappresentano ancora una valutazione preliminare della qualità dei modelli, poiché ottenuti su dati di validazione derivanti dallo stesso dominio del training, seppur modificati artificialmente. La reale capacità di generalizzazione sarà oggetto di approfondimento nella successiva fase di validazione esterna.

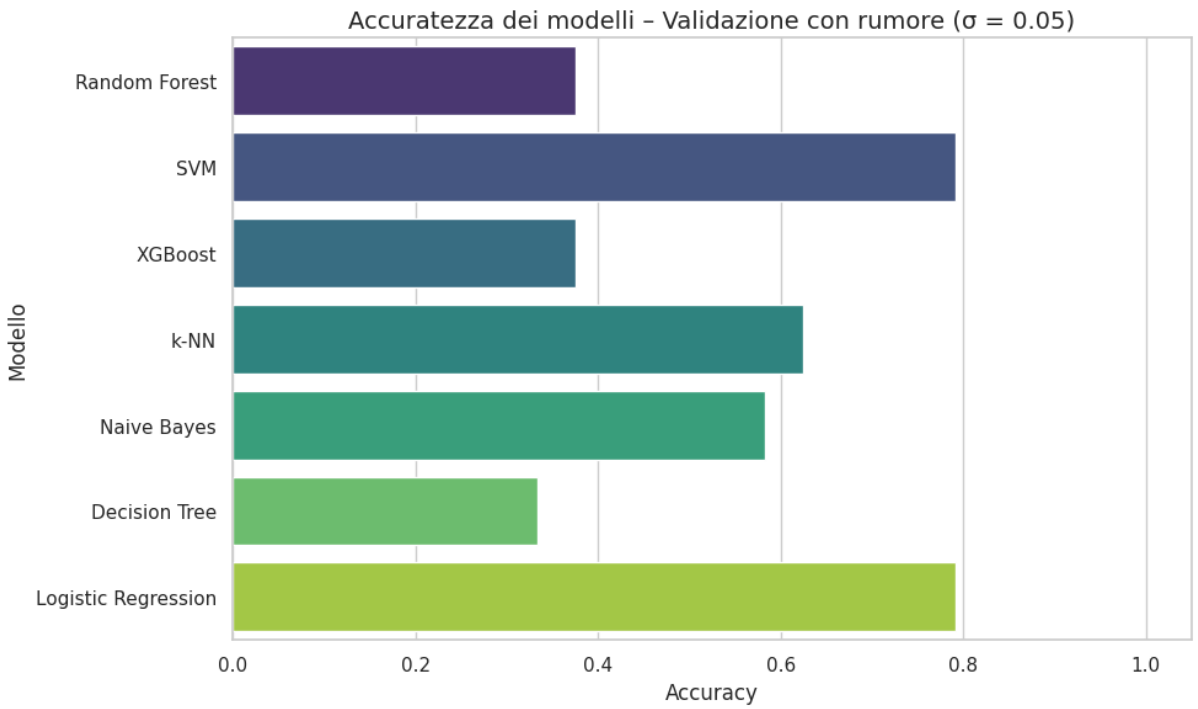


Figura 4.28 - Concentrato di Pomodoro - Accuracy per ciascun modello durante predizione su dati non visti.

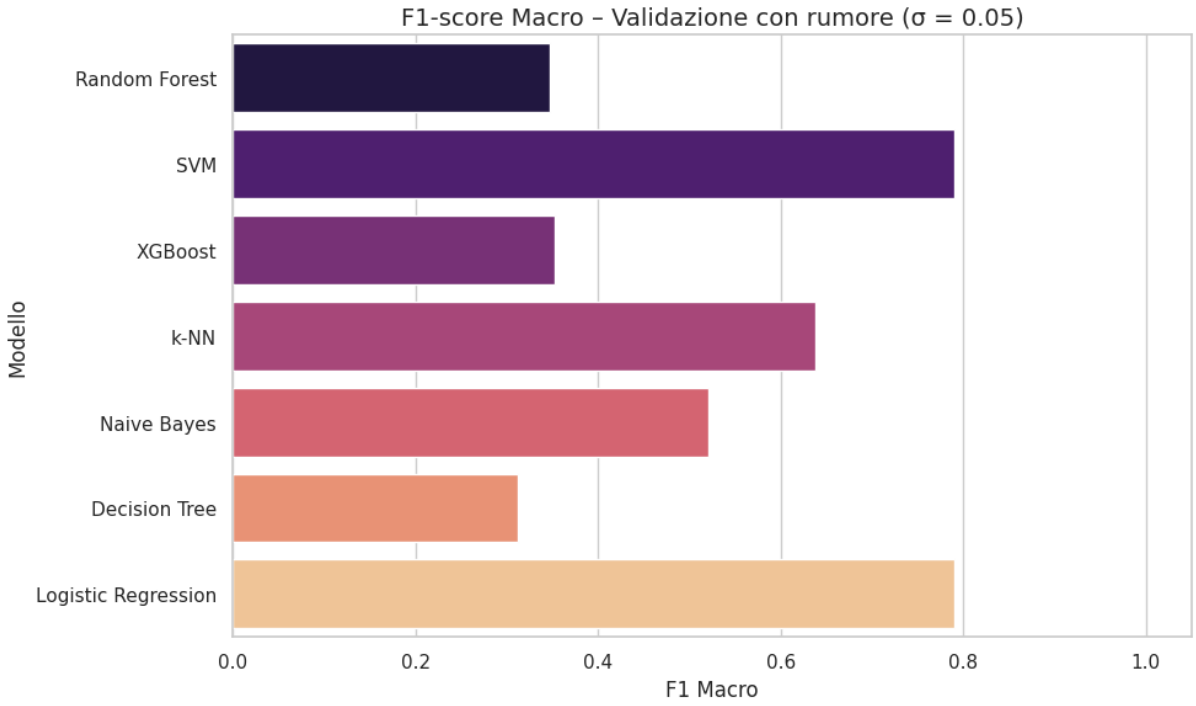


Figura 4.29 - Concentrato di Pomodoro - F1-score Macro per ciascun modello durante predizione su dati non visti

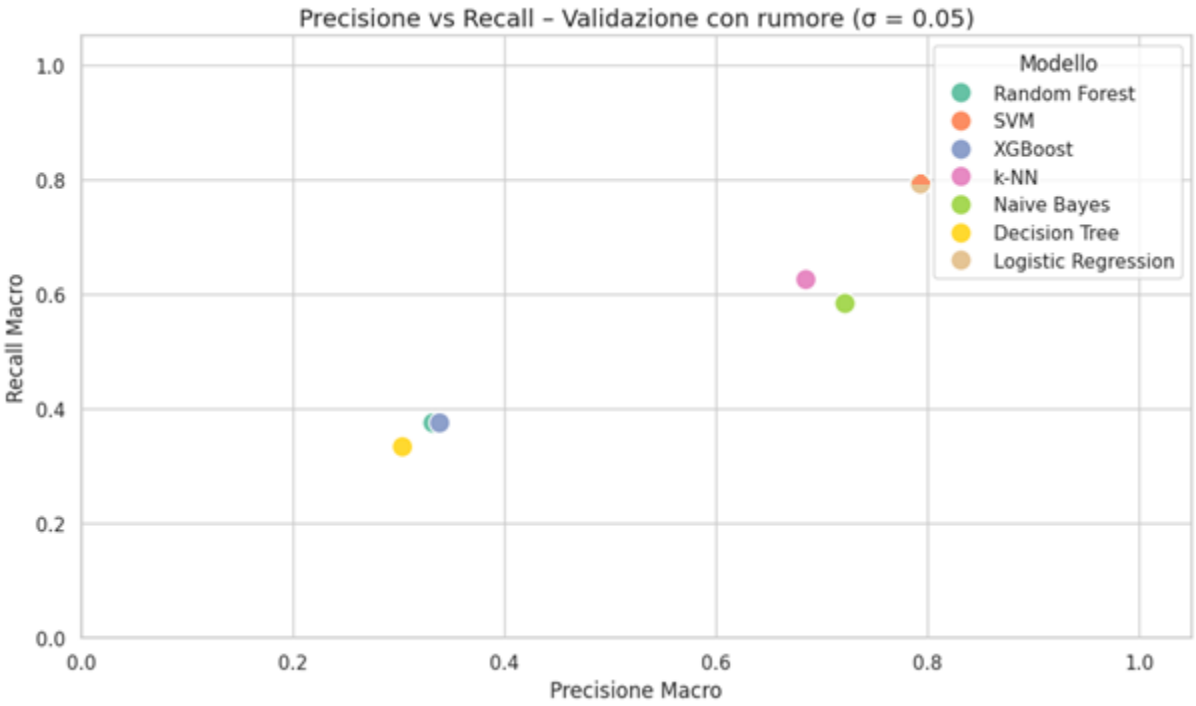


Figura 4.30 - Concentrato di Pomodoro - Precisione vs Recall macro per ciascun modello durante predizione su dati non visti.

Classificazione finale e selezione delle feature più rilevanti

I risultati riportati nella Tabella 4.14 evidenziano le prime 20 variabili (buckets NMR) che hanno mostrato la maggiore rilevanza durante il processo di classificazione condotto dai diversi algoritmi di machine learning, valutata mediante il criterio della *Mean Decrease Accuracy* (MDA). L'analisi si riferisce alla fase di test dei modelli eseguita su dati affetti da rumore gaussiano di intensità moderata ($\sigma = 0.1$), al fine di simulare condizioni realistiche di variabilità sperimentale.

Le variabili sono elencate in ordine decrescente rispetto al valore massimo di importanza MDA registrato, e sono accompagnate dagli intervalli spettrali corrispondenti (espressi in ppm) e dal numero di modelli predittivi nei quali sono risultate significative per la classificazione.

Tabella 4.14 – Concentrato di Pomodoro - Principali variabili (buckets NMR) selezionate tramite MDA durante il processo di classificazione.

Num	Variabile	PPM		Totale modelli	Importanza MDA
1	Var97	5.74	±0.02	6	3.403
2	Var99	5.66	±0.02	6	3.276
3	Var98	5.70	±0.02	5	3.216
4	Var96	5.78	±0.02	5	3.004
5	Var100	5.62	±0.02	6	2.972
6	Var4	9.46	±0.02	7	2.959
7	Var94	5.86	±0.02	3	2.831
8	Var102	5.54	±0.02	6	2.809
9	Var101	5.58	±0.02	4	2.705
10	Var95	5.82	±0.02	1	2.616
11	Var90	6.02	±0.02	1	2.527
12	Var85	6.22	±0.02	1	2.406
13	Var82	6.34	±0.02	2	2.382
14	Var10	9.22	±0.02	5	2.330
15	Var103	5.50	±0.02	1	2.288
16	Var88	6.10	±0.02	1	2.285
17	Var83	6.30	±0.02	2	2.273
18	Var77	6.54	±0.02	3	2.252
19	Var78	6.50	±0.02	4	2.212
20	Var84	6.26	±0.02	1	2.209

Le prime 20 variabili selezionate mostrano valori di importanza compresi tra 2.209% e 3.403%, evidenziando un contributo rilevante e differenziato al processo di classificazione. L'analisi incrociata tra l'entità dell'importanza (MDA) e il numero di modelli nei quali ciascuna variabile consente inoltre di identificare i segnali più robusti e ricorrenti.

Tra le variabili con maggiore rilevanza predittiva si distinguono Var97 (5.74 ppm), Var99 (5.66 ppm), Var98 (5.70 ppm), Var96 (5.78 ppm) e Var100 (5.62 ppm), tutte caratterizzate da MDA superiori al 2.9% e selezionate da cinque o sei modelli. Questo gruppo di segnali, localizzati in una ristretta regione spettrale tra 5.62 e 5.78 ppm, risulta particolarmente coerente e stabile nel supportare la classificazione.

Di rilievo anche Var4 (9.46 ppm), che presenta un valore MDA pari a 2.959% ed è l'unica variabile selezionata da tutti e sette i modelli, a conferma del suo impatto costante nel processo predittivo.

All'opposto, variabili come Var95 (5.82 ppm), Var90 (6.02 ppm), Var85 (6.22 ppm), Var103 (5.50 ppm) e Var84 (6.26 ppm), pur mostrando valori di MDA superiori a 2.2%, sono state selezionate da un solo modello. Ciò suggerisce una minore consistenza trasversale e invita a una valutazione più cauta del loro ruolo discriminante.

Validazione mediante set di dati esterno

Nella tabella seguente è riportata la performance dei diversi modelli in termini di accuratezza nella predizione su un set di dati esterno, privo di rumore gaussiano, ottenuto come descritto nel paragrafo 3.2.2.1.

In Figura 4.31 viene presentata, a titolo esemplificativo, la matrice di confusione relativa al modello Random Forest ottenuta nella fase di predizione sul set esterno.

Tabella 4.15 - Concentrato di Pomodoro - Accuratezza, F1-score macro e F1-score weighted dei modelli di classificazione applicati al set esterno di dati

Modello	Accuracy	F1_macro	F1_weighted
Random Forest	1	1	1
SVM	1	1	1
XGBoost	1	1	1
k-NN	0.708	0.698	0.698
Naive Bayes	0.917	0.915	0.915
Decision Tree	1	1	1
Logistic Regression	1	1	1

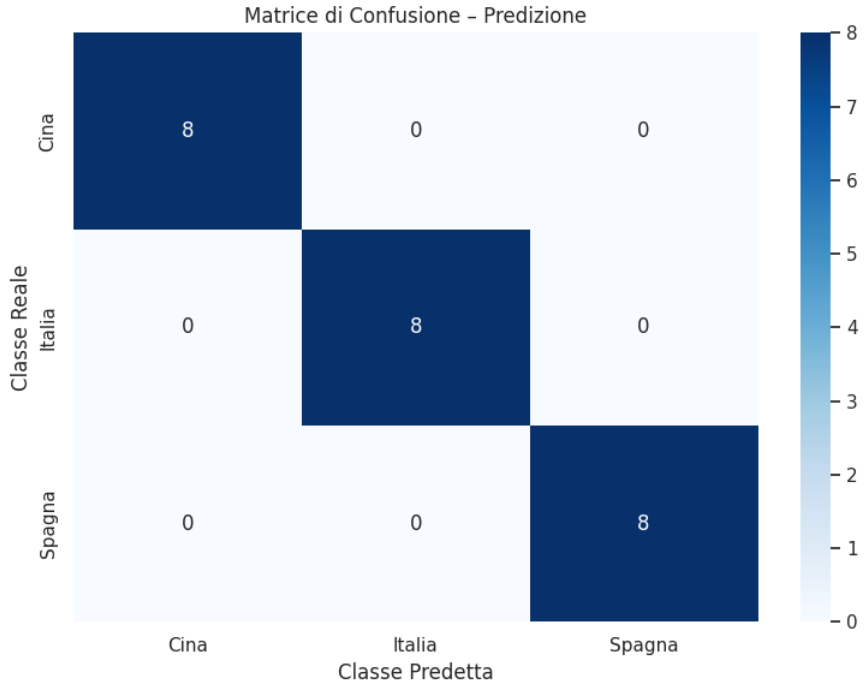


Figura 4.31 - Concentrato di Pomodoro - Matrice di confusione ottenuta con il modello Random Forest durante classificazione con set esterno di dati.

I risultati ottenuti confermano un'elevata capacità predittiva per la maggior parte degli algoritmi testati: Logistic Regression, SVM, Random Forest, XGBoost e Decision Tree raggiungono un'accuratezza del 100%, accompagnata da valori perfetti anche per F1-score macro e weighted, a testimonianza della loro capacità di generalizzare in modo impeccabile su dati non visti. Il modello Naive Bayes mostra una leggera diminuzione delle prestazioni, con un'accuratezza pari al 91,7% e valori di F1 macro e weighted ugualmente elevati (0.915), suggerendo un'ottima capacità di classificazione, seppur con qualche marginale incertezza. Più contenute risultano invece le prestazioni del modello k-NN, che registra un'accuratezza del 70,8%, penalizzato probabilmente dalla natura locale del metodo e dalla sua maggiore sensibilità alla distribuzione spaziale dei punti nel nuovo spazio proiettato.

La matrice di confusione per Random Forest evidenzia una perfetta classificazione di tutti i campioni appartenenti alle tre classi (Cina, Italia e Spagna), confermando l'efficacia del modello nel discriminare correttamente l'origine geografica del concentrato di pomodoro in condizioni operative realistiche. I risultati ottenuti sul set esterno dimostrano che i modelli addestrati risultano particolarmente robusti e affidabili nella fase di predizione reale.

Discussione finale

L'approccio automatizzato basato sull'integrazione tra PCA e modelli di machine learning supervisionato ha mostrato alcune criticità durante la discriminazione tra campioni di Concentrato di Pomodoro provenienti da diverse aree geografiche.

I risultati ottenuti durante la fase di addestramento con dati affetti da rumore gaussiano (σ = 0.1) sono stati nel complesso modesti. Solo Logistic Regression e SVM hanno raggiunto accuratezze superiori al 75%, mentre altri modelli, come Random Forest, Decision Tree e XGBoost, hanno mostrato performance limitate, con accuratezze vicine al livello casuale. Questa eterogeneità evidenzia la difficoltà degli algoritmi nel discriminare efficacemente tra le classi in condizioni di rumore, probabilmente a causa della natura complessa e poco marcata delle differenze tra i profili metabolici.

La selezione delle variabili mediante MDA ha comunque permesso di individuare alcune regioni spettrali di rilievo, in particolare tra 5.62 e 5.78 ppm, con segnali ricorrenti e importanza elevata. Var4 (9.46 ppm), selezionata da tutti i modelli, si distingue per coerenza e rilevanza, suggerendo un potenziale ruolo chiave nella classificazione.

La fase di validazione esterna, condotta su un set indipendente privo di rumore, ha ribaltato parzialmente le indicazioni ottenute durante l'addestramento. I modelli Logistic Regression, SVM, Random Forest, XGBoost e Decision Tree hanno raggiunto performance eccellenti, con accuratezze pari al 100% e F1-score perfetti. Ciò indica che, nonostante la difficoltà nel gestire dati perturbati, i modelli sono stati in grado di apprendere pattern informativi generalizzabili. Le prestazioni inferiori del k-NN confermano la sensibilità del metodo alla struttura locale dei dati. In sintesi, pur in presenza di risultati iniziali poco soddisfacenti in fase di addestramento, il workflow AI-driven ha dimostrato una discreta efficacia durante la classificazione su dati reali non affetti da rumore.

4.3.4 Confronto tra approccio classico e automatizzato

Il confronto tra l'approccio chemiometrico convenzionale e quello automatizzato AI-driven applicato ai campioni di concentrato di pomodoro evidenzia differenze sostanziali in termini di efficacia analitica, stabilità predittiva e profondità interpretativa, pur rivelando alcuni elementi di convergenza tra le due strategie. Dal punto di vista esplorativo, l'analisi PCA condotta con metodo classico ha consentito di rilevare una parziale separazione tra le classi geografiche (Italia, Spagna, Cina), con una distribuzione spaziale che riflette differenze nella varianza intraclasse dei campioni. La selezione delle variabili discriminanti, basata sull'analisi dei contributi alle componenti principali e confermata tramite ANOVA e test di Tukey, ha permesso

di individuare numerosi segnali significativi, in particolare nella regione degli zuccheri (δ 3.2–5.4 ppm), degli acidi organici (δ 1.5–2.0 ppm) e dei composti aromatici (δ 7.5–8.5 ppm). L'assegnazione spettrale ha confermato il ruolo centrale di metaboliti quali glucosio, fruttosio, GABA, acido glutammico e trigonellina, già noti in letteratura per la loro rilevanza nella caratterizzazione chimica del pomodoro e dei suoi derivati industriali (Bénard et al., 2015; Ingallina et al., 2020; Pin et al., 2025). L'approccio automatizzato, basato sull'integrazione tra PCA, modelli di machine learning supervisionato e selezione delle feature tramite MDA, ha mostrato un comportamento più complesso. Durante la fase di addestramento su dati affetti da rumore gaussiano ($\sigma = 0.1$), le performance predittive sono risultate eterogenee e, in alcuni casi, inferiori a quanto atteso. Solo Logistic Regression e SVM hanno raggiunto accuratezze superiori al 75%, mentre modelli come Random Forest, XGBoost e Decision Tree hanno mostrato prestazioni vicine al livello casuale, suggerendo una bassa resilienza ai dati perturbati per questa specifica matrice. Tuttavia, la validazione esterna ha evidenziato un'inversione di tendenza: gli stessi modelli che in fase di training risultavano deboli hanno ottenuto accuratezze perfette (1.000), dimostrando un'elevata capacità di generalizzazione in condizioni operative realistiche. Tale risultato conferma che i pattern informativi, seppur deboli e mascherati dal rumore, risultano effettivamente codificabili dai modelli quando applicati a dati non perturbati. Un elemento di convergenza tra i due approcci riguarda l'individuazione di segnali rilevanti nella regione δ 5.62–5.78 ppm (es. Var96–Var100), altamente informativi nel processo di classificazione AI-driven e riconducibili a protoni anomerici di zuccheri e pectine. Anche la variabile Var4 (δ 9.46 ppm), assegnata alla trigonellina, è emersa come discriminante in entrambi i metodi, confermando la rilevanza di questo composto nella differenziazione tra i campioni. Sebbene i criteri di selezione delle variabili siano diversi – statistici e orientati alla spiegazione nel metodo classico, predittivi e orientati alla performance nell'approccio AI – la sovrapposizione di alcuni segnali chiave suggerisce una parziale coerenza tra le due strategie analitiche. In sintesi, l'approccio classico ha fornito un'interpretazione più lineare e diretta della variabilità metabolomica, utile per una caratterizzazione qualitativa e biochimica del prodotto. L'approccio automatizzato, pur mostrando limiti in condizioni perturbate, ha confermato una buona efficacia in fase predittiva su dati reali e ha permesso di identificare segnali stabili e ricorrenti attraverso modelli indipendenti. La convergenza su variabili comuni (es. trigonellina, zuccheri, GABA) rappresenta un punto di forza metodologico, rafforzando l'affidabilità delle conclusioni emerse. I due approcci risultano complementari: il metodo classico utile per una comprensione biochimica strutturata, l'AI-driven per la costruzione di strumenti predittivi scalabili e applicabili in contesti industriali.

4.3.5 Esempio di serializzazione e generazione hash per registrazione del fingerprint metabolico su blockchain

Di seguito viene mostrato un esempio di file serializzato in formato JSON ottenuto a partire dal segnale NMR grezzo (FID) acquisito da un campione rappresentativo della matrice Concentrato di Pomodoro (codice Romano_Con4_r3), attraverso spettroscopia 1H HR-MAS.


```
{
  "sample_id": "Romano_Con4_r3",
  "matrix": "Concentrati Pomodoro",
  "instrument": "av400",
  "operator": "Cangemi",
  "date": "2021-10-18T14:32:53",
  "temperature": 298.5393,
  "pulse_program": "zgpr",
  "fid_real": [
    0.0,
    0.0,
…
    451.0
  ],
  "fid_imag": [
    0.0,
    0.0,
...
    511.0
  ],
  "is_domain": "time",
  "timestamp": "2025-06-10T06:28:13.000673"
}
```


Figura 4.32 – Concentrato di Pomodoro - Esempio di serializzazione da FID Bruker

Il file JSON include la parte reale e quella immaginaria del FID e un insieme di metadati contestuali: strumento, operatore, data, temperatura, sequenza di impulso e timestamp.

Il contenuto del file JSON, sottoposto a funzione di hash SHA256, ha fornito il seguente identificativo digitale:

08107712513cffb3c9407763761accee7957ba9f7068c46f7822d78b9b215a4d

Figura 4.33 – Concentrato di Pomodoro - Esempio di generazione hash da FID Bruker

L’hash così ottenuto può fungere da identificativo digitale non modificabile, utile per ancorare il dato analitico a uno smart contract distribuito.
Questo hash rappresenta una "firma digitale" univoca del file, che ne garantisce l’integrità e l’autenticità. Una volta registrato su una rete blockchain (es. Hyperledger Fabric), tale impronta permette di verificare, in qualsiasi momento, che il file non sia stato modificato e che l’acquisizione sia avvenuta in condizioni sperimentali note e tracciabili.

4.4 Latte UHT

4.4.1 Descrizione dei profili spettrali

A titolo esemplificativo, la Figura 4.34 mostra gli spettri 1H-NMR relativi a tre campioni di latte UHT appartenenti allo stesso marchio commerciale (Arborea, Italia), rappresentativi delle principali categorie differenziate in base al contenuto lipidico: scremato (0.1%), parzialmente scremato (1.55%) e intero (3.6%).

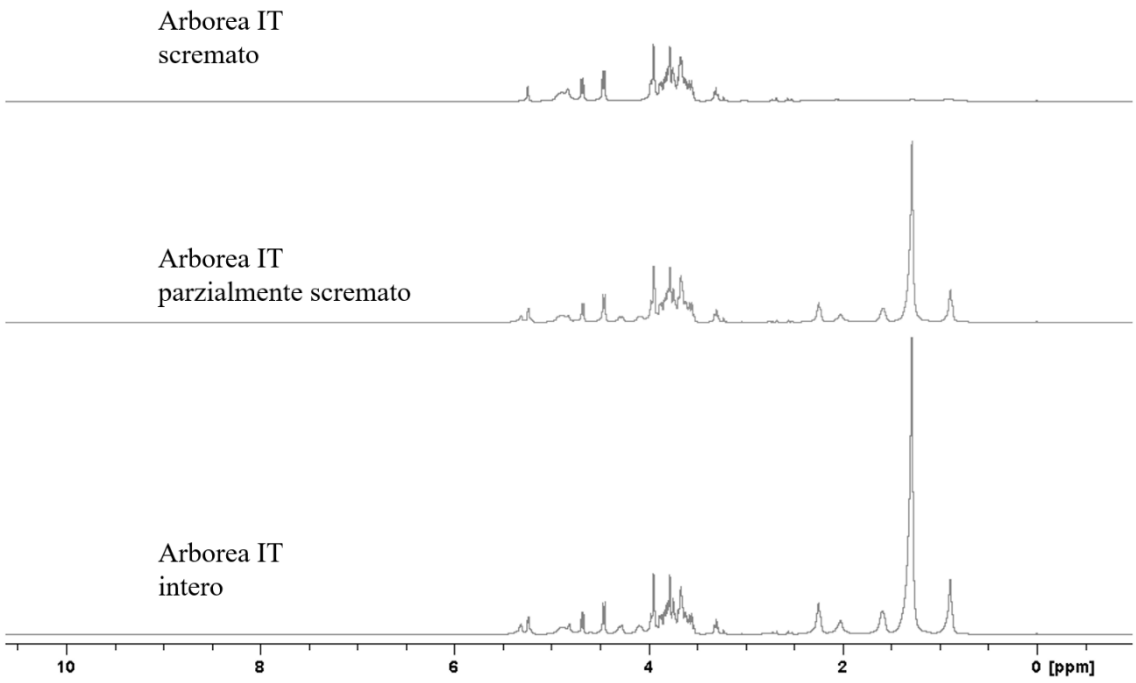


Figura 4.34 – Latte UHT - Spettri ^{1}H HR-MAS NMR relativi a tre tipologie della marca Arborea (Italia): intero, parzialmente scremato e scremato.

L'effetto della variazione della frazione grassa sul profilo spettrale risulta evidente: si osserva infatti un progressivo decremento dell'intensità dei segnali caratteristici dei lipidi, in particolare quelli associati ai gruppi metilici (-CH_3, ~0.9 ppm) e metilenici (-CH_2-, ~1.2–1.4 ppm), passando dal latte intero a quello scremato. Le regioni comprese tra 3.0 e 4.5 ppm, riconducibili principalmente a zuccheri quali lattosio (~3.65–4.10 ppm), glucosio (~3.20–3.90 ppm), galattosio (~3.40–4.00 ppm) e relativi derivati come l'N-acetilglucosammina (~2.00–2.10 ppm per il gruppo acetilico, ~3.40–4.00 ppm per la parte zuccherina), non mostrano variazioni rilevanti tra le diverse tipologie di latte. Allo stesso modo, le bande attribuibili a colina (~3.20 ppm), creatina (~3.05 ppm) e citrato (~2.65 ppm) risultano pressoché invariate, suggerendo una sostanziale stabilità di questi metaboliti indipendentemente dal tenore lipidico (Eltemur et al., 2023; Mazzei & Piccolo, 2018; Tsermoula et al., 2024). Infine, la regione aromatica compresa

tra 6.5 e 9.0 ppm presenta segnali molto deboli, verosimilmente associabili ad amminoacidi aromatici, come l'istidina (~7.10–8.10 ppm) e la tirosina (~6.90–7.20 ppm), nonché a nucleotidi quali l'adenina (~8.20 ppm) e l'uridina (~7.80 ppm), in linea con quanto descritto in letteratura (Sundekilde et al., 2013).

4.4.2 Risultati dell'analisi chemiometrica convenzionale

PCA

La Figura 4.35 illustra la distribuzione dei campioni di latte UHT nello spazio delle componenti principali, rappresentate nei piani F2–F5 (a sinistra) e F2–F3–F5 (a destra), così come ottenuto dall'analisi condotta con XLSTAT.

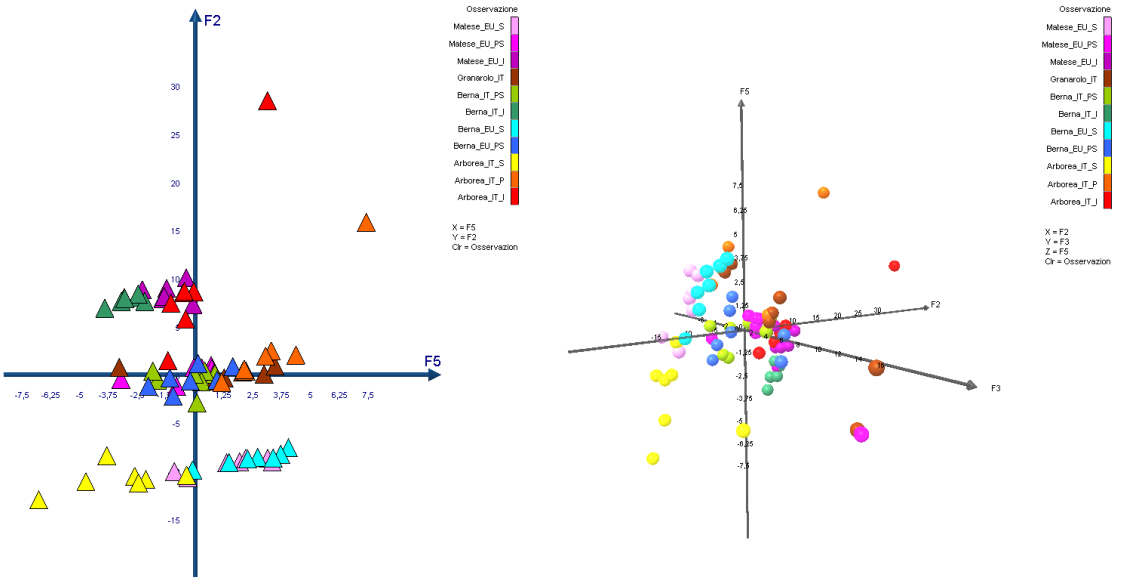


Figura 4.35 – Latte UHT - Score Plot da PCA.

La separazione osservata lungo l'asse F2 nel grafico bidimensionale evidenzia una chiara distinzione basata sul contenuto lipidico: i campioni di latte intero si collocano nella parte superiore, quelli scremati nella porzione inferiore, mentre i parzialmente scremati si distribuiscono in posizione intermedia. Questo andamento suggerisce che la componente principale F2 cattura gran parte della variabilità associata alla frazione grassa.

Nel piano F5, invece, non si osservano pattern altrettanto marcati, sebbene siano visibili alcune tendenze di separazione riconducibili ai diversi marchi commerciali. La rappresentazione tridimensionale (F2–F3–F5), sebbene consenta una lettura più articolata della distribuzione,

non evidenzia tendenze nette riconducibili a fattori di origine geografica o trattamento industriale. Tale risultato appare coerente con l'elevato livello di standardizzazione del latte UHT a livello europeo, che tende a uniformare le differenze legate a variabili produttive secondarie.

Selezione delle variabili significative

La Tabella 4.16 riporta, per ciascuna delle prime cinque componenti principali, l'elenco delle variabili che risultano significativamente differenti tra almeno due gruppi ($p < 0.05$), sulla base dell'analisi della varianza univariata (ANOVA). È opportuno precisare che la significatività statistica indicata non implica che ciascuna variabile discrimini tutte le classi tra loro, bensì che, per almeno un confronto tra coppie di gruppi, è stata rilevata una differenza significativa.
Per una maggiore chiarezza interpretativa, è stata successivamente condotta un'analisi post-hoc (test di Tukey HSD) al fine di individuare con precisione quali gruppi differiscono tra loro per ciascuna variabile significativa. Nella presente tabella è tuttavia riportato esclusivamente l'elenco complessivo delle variabili che hanno mostrato una significatività globale (ANOVA a $p < 0.05$), senza entrare nel dettaglio dei confronti specifici tra classi. Tale approccio è stato adottato con l'obiettivo di fornire una sintesi preliminare delle variabili maggiormente coinvolte nella separazione dei gruppi lungo ciascuna componente principale.

Tabella 4.16 – Latte UHT - Elenco delle variabili significative.

	Selezione PCA+ANOVA+Tukey
PC	Variabili significative ($p < 0.05$)
PC1	Var48, Var50, Var51, Var52, Var53, Var57, Var58, Var59, Var60, Var61, Var62, Var63, Var64, Var70, Var71, Var72, Var73
PC2	Var107, Var108, Var128, Var129, Var133, Var179, Var180, Var185, Var195, Var196, Var199, Var202, Var203, Var207, Var213
PC3	Var111, Var125, Var148, Var149, Var150, Var154, Var155, Var157, Var158, Var181, Var205
PC4	Var111, Var125, Var148, Var150, Var154, Var157, Var158, Var159, Var164, Var165, Var175, Var181, Var205
PC5	Var1, Var2, Var3, Var4, Var5, Var6, Var7, Var8, Var10, Var11, Var136

Assegnazione delle variabili discriminanti ai metaboliti

Nella tabella seguente è riportata l'assegnazione delle principali variabili (bucket spettrali) ai metaboliti putativi, effettuata mediante confronto con la libreria metabolomica precedentemente predisposta.

Tabella 4.17 – Latte UHT - Assegnazione putativa delle variabili ai metaboliti.

Buckets	PC	δ ^{1}H (ppm)	δ ^{13}C (ppm)	Metabolita
Var57	PC1	7,34		Fenilalanina
Var58	PC1	7,3		Fenilalanina
Var63	PC1	7,1		Istidina
Var64	PC1	7,06		Istidina
Var202	PC2	1,34	20–75; 20–75; 55–65	Acido lattico; Lattato; Treonina
Var213	PC2	0,9	14–35	Acido butirrico
Var148	PC3	3,5	70–75; 60–105	Inositolo; Glucosio
Var154	PC3	3,26	60–105	Glucosio
Var155	PC3	3,22	54–56; 60–105	Colina; Glucosio
Var205	PC3	1,22	24–65; 18–60	Beta-idrossibutirrato; Etanolo
Var148	PC4	3,5	70–75; 60–105	Inositolo; Glucosio
Var150	PC4	3,42	55–70; 60–105	Carnitina; Glucosio
Var154	PC4	3,26	60–105	Glucosio
Var158	PC4	3,1	40–60	Lisina
Var159	PC4	3,06	45–50; 40–60	Creatina; Lisina
Var165	PC4	2,82	35–55	Aspartato
Var175	PC4	2,42	30–35	Acido succinico
Var205	PC4	1,22	24–65; 18–60	Beta-idrossibutirrato; Etanolo

PC1 ha evidenziato segnali aromatici localizzati a 7.34 e 7.30 ppm (Var57, Var58), riconducibili alla fenilalanina, e a 7.10 e 7.06 ppm (Var63, Var64), compatibili con la presenza di istidina, entrambi amminoacidi aromatici comunemente rilevati nei profili NMR del latte. Lungo PC2, è stata osservata una variabile significativa a 1.34 ppm (Var202), che mostra una sovrapposizione tra le regioni tipiche di acido lattico, lattato e treonina, tutti composti correlati al metabolismo energetico e alla fermentazione. A 0.90 ppm (Var213), invece, è stato identificato un segnale caratteristico dell'acido butirrico, un acido grasso a corta catena tipico del latte ruminante e indicativo della fermentazione batterica. Per PC3, è emersa la presenza di segnali nella regione 3.50–3.22 ppm (Var148, Var154, Var155), coerenti con la presenza di inositolo, glucosio e colina, metaboliti frequentemente riscontrati nei latticini e coinvolti in processi osmoregolatori e di trasporto di gruppi metilici. La variabile Var205 (1.22 ppm) si è dimostrata invece riconducibile a beta-idrossibutirrato ed etanolo, entrambi indicatori di fermentazione o alterazione metabolica. Anche PC4 ha confermato e approfondito il contributo

dei segnali zuccherini e osmoliti (Var148, Var150, Var154), ma ha inoltre evidenziato segnali a 3.10–3.06 ppm (Var158, Var159), compatibili con lisina e creatina, due composti chiave per l'attività muscolare e il metabolismo azotato. Infine, la presenza di aspartato (Var165, 2.82 ppm) e acido succinico (Var175, 2.42 ppm) fornisce ulteriori evidenze di attività metabolica legata al ciclo degli acidi tricarbossilici (TCA).

4.4.3 Risultati dell'analisi chemiometrica AI-driven

Importazione, verifica e pre-processing dei dati

La Figura 4.36 mostra un estratto della matrice dei bucket spettrali importata nello script, rappresentando i dati prima e dopo la standardizzazione mediante trasformazione Z-score.
Tale standardizzazione realizza una distribuzione centrata attorno allo zero e priva di dominanze numeriche, condizione ottimale per l'applicazione di modelli di machine learning supervisionato. Questo passaggio garantisce che ogni variabile contribuisca in egual misura al modello, prevenendo distorsioni dovute alla diversa scala delle intensità.

Dati Grezzi (prime 5 righe):

	replica	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	...	Var222
0	1	75038.13335	82856.98555	85414.72181	78847.33473	86941.68992	105912.8606	107995.2456	111276.7632	108343.0538	...	1290620.867
1	2	191074.99720	155568.95090	181064.70470	218725.13950	264673.96870	253416.2439	240489.5716	281350.3306	313941.1691	...	8078546.553
2	3	132043.14510	134058.86550	165401.38100	179612.07870	184569.24860	186369.6873	169623.9763	186620.3816	185875.6073	...	3608146.583
3	4	149452.30390	169886.33620	178473.92850	187934.64210	206840.13510	209627.0924	188727.6678	231744.3184	226910.8587	...	3700589.426
4	5	123703.43050	138802.14920	150900.17210	172955.94860	165920.30150	175024.8094	182882.5838	191430.6296	200405.3920	...	3480513.725

5 rows × 232 columns

Dati Standardizzati (prime 5 righe):

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224
0	-1.142567	-1.118509	-1.212595	-1.366696	-1.407195	-1.271052	-1.248890	-1.322262	-1.357635	-1.381538	...	-2.300225	-2.327080	-2.346044	-2.326889
1	0.130599	-0.370742	-0.278047	-0.124582	0.090459	-0.128296	-0.240016	-0.010036	0.179835	0.286857	...	3.826588	3.920091	3.945750	3.921237
2	-0.517103	-0.591951	-0.431086	-0.471906	-0.584540	-0.647727	-0.779619	-0.740938	-0.777844	-0.640021	...	-0.188256	-0.194178	-0.181710	-0.197231
3	-0.326088	-0.223503	-0.303361	-0.398002	-0.396875	-0.467545	-0.634155	-0.392778	-0.470981	-0.462793	...	-0.186080	-0.109100	-0.110305	-0.115463
4	-0.608607	-0.543171	-0.572770	-0.531012	-0.741685	-0.735620	-0.678662	-0.703824	-0.669189	-0.701909	...	-0.307707	-0.311643	-0.279230	-0.277835

5 rows × 230 columns

Figura 4.36 – Latte UHT - Estratto della matrice dei bucket spettrali prima e dopo la standardizzazione Z-score

Riduzione dimensionale mediante PCA

In analogia con quanto mostrato nel paragrafo 4.4.2, vengono presentati i risultati dell'analisi PCA condotta con l'impiego dello script. La Figura 4.37 riporta l'andamento della varianza cumulativa spiegata (Elbow Plot), evidenziando come la soglia del 95% venga raggiunta entro le prime cinque componenti principali. La combinazione dei criteri adottati — soglia del 95% della varianza spiegata, criterio di Kaiser e metodo del "Broken Stick" — ha condotto alla selezione delle prime cinque componenti principali, ritenute ottimali per l'addestramento e la predizione dei modelli di machine learning.

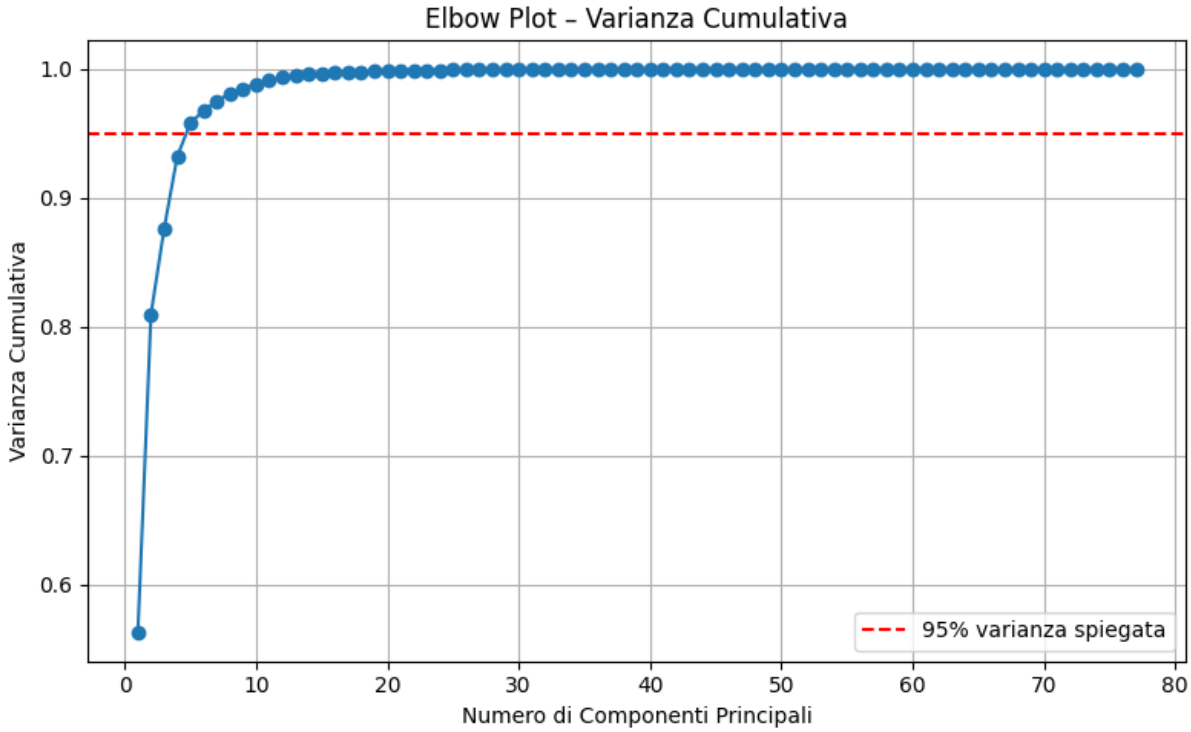


Figura 4.37 – Latte UHT - Elbow Plot relativo alla varianza cumulativa spiegata dalle componenti principali e indicazione della soglia al 95%.

Gli score plot bidimensionali ottenuti dall'analisi PCA sono rappresentati graficamente in Figura 4.38.

L'esame delle combinazioni biplot evidenzia una certa separazione tra i gruppi, in particolare in relazione al contenuto lipidico, ma non mostra una netta distinzione sistematica tra tutti i marchi.

Dall'osservazione complessiva emerge che la separazione più evidente si manifesta lungo la componente PC2, particolarmente nei grafici PC1 vs PC2, PC2 vs PC4 e PC2 vs PC5. In queste proiezioni, è possibile distinguere un chiaro gradiente verticale che riflette la diversa concentrazione della frazione lipidica nei campioni: i campioni di latte intero tendono a posizionarsi nelle regioni superiori del piano PC2, mentre quelli di latte scremato occupano la parte inferiore, con i campioni di latte parzialmente scremato che si collocano in posizione intermedia. Questo andamento suggerisce che la componente PC2 rappresenti la principale fonte di variabilità legata al contenuto lipidico, e risulta coerente con quanto osservato nelle analisi qualitative degli spettri

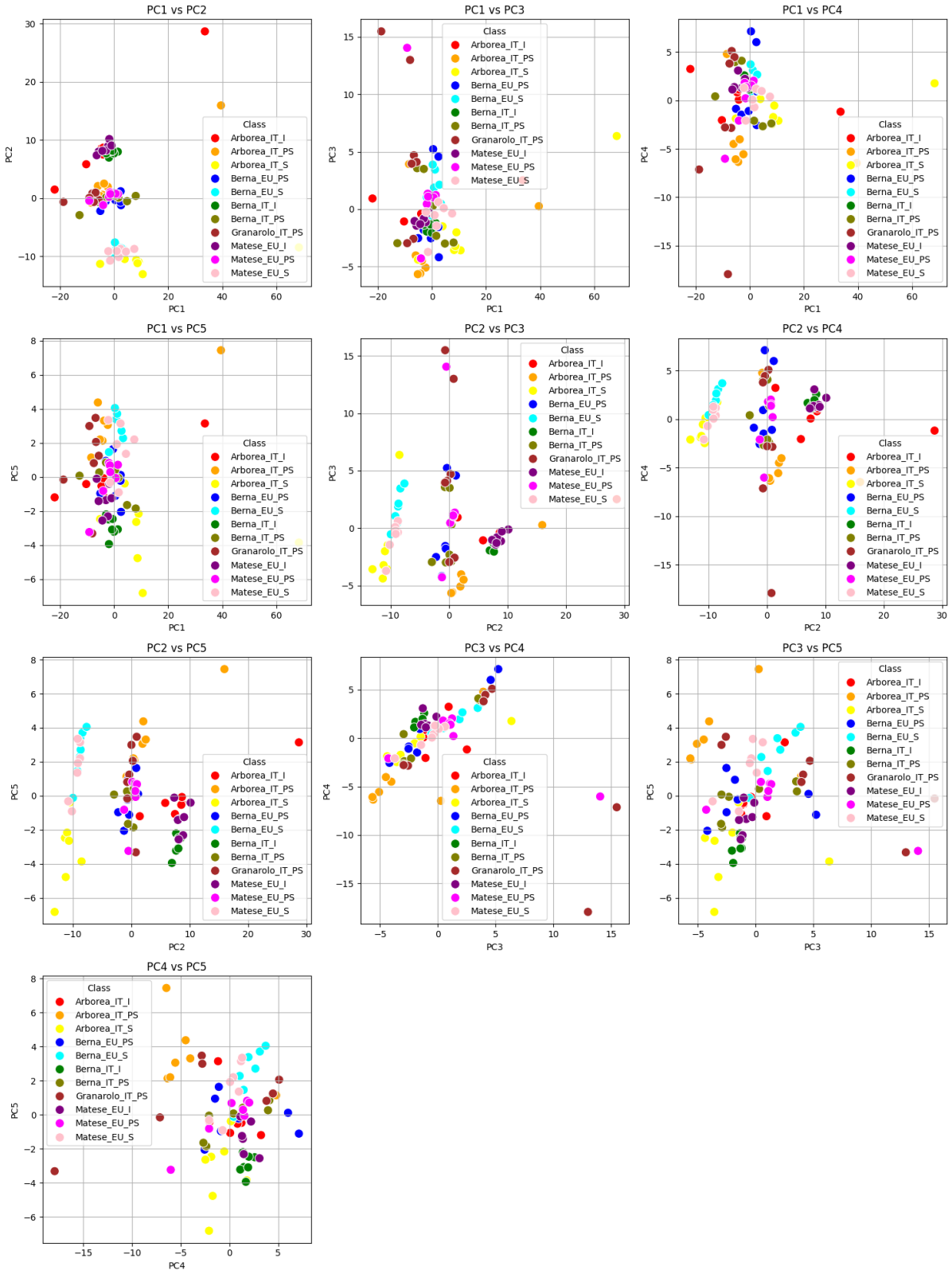


Figura 4.38 – Latte UHT - Distribuzione delle osservazioni nello spazio bidimensionale considerando la combinazione delle prime 5 PC.

La componente PC1, pur rappresentando una quota significativa della varianza totale, non sembra correlata in modo evidente a un singolo fattore discriminante, come confermato dalla parziale sovrapposizione delle classi nei grafici PC1 vs PC3 e PC1 vs PC5. Anche le componenti PC3, PC4 e PC5 non evidenziano tendenze chiare di separazione tra i gruppi, sebbene in alcuni piani (es. PC3 vs PC4) si notino dispersioni localizzate che potrebbero riflettere differenze minori legate a variabili produttive secondarie, come l'origine geografica o la linea di produzione.

Nel complesso, l'analisi PCA evidenzia una separazione parziale dei campioni basata sul contenuto di grassi, ma non permette una discriminazione netta tra tutte le classi presenti, soprattutto nei confronti dei diversi marchi. Questo risultato è coerente con l'elevato grado di omogeneizzazione che caratterizza il latte UHT, un prodotto fortemente standardizzato a livello industriale. Tale standardizzazione comporta una ridotta variabilità compositiva, anche tra campioni di diversa origine o appartenenti a marchi differenti, limitando così l'efficacia delle tecniche esplorative non supervisionate nel rilevare differenze significative.

Addestramento dei modelli di machine learning supervisionato

I risultati riportati nella Tabella 4.18 illustrano le performance predittive ottenute durante la fase di addestramento dei modelli di classificazione. L'analisi è stata condotta mediante validazione incrociata stratificata a 5 fold, in presenza di rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza dei modelli rispetto a una variabilità simulata. Per ciascun algoritmo sono riportate l'accuratezza media, la deviazione standard e gli iperparametri ottimali individuati tramite ottimizzazione con GridSearchCV.

Tabella 4.18 – Latte UHT - Prestazioni predittive dei modelli di classificazione durante l'addestramento con dati affetti da rumore gaussiano

Modello	Accuracy (mean)	Accuracy (std)	Parametri ottimali
Random Forest	0,571	0,094	{'max_depth': None, 'n_estimators': 200}
SVM	0,584	0,136	{'C': 1, 'kernel': 'linear'}
XGBoost	0,533	0,090	{'learning_rate': 0.01, 'max_depth': 6
k-NN	0,597	0,095	{'metric': 'euclidean', 'n_neighbors': 3}
Naive Bayes	0,416	0,066	default
Decision Tree	0,558	0,123	{'max_depth': None}
Logistic Regression	0,584	0,095	{'C': 1}

Nel complesso, le performance ottenute risultano moderate, con valori di accuratezza mediamente compresi tra 0.41 e 0.60. Questo risultato è attribuibile sia alla natura fortemente

standardizzata della matrice analizzata, sia al fatto che i modelli sono stati addestrati su un set di dati artificialmente perturbato mediante l'aggiunta di rumore gaussiano, in una fase intermedia del workflow. Le prestazioni effettive verranno successivamente valutate nella fase di classificazione su set esterni, non affetti da rumore, al fine di stimare la reale capacità di generalizzazione dei modelli.

Nei grafici seguenti sono riportate le metriche di performance dei modelli ottenute durante la fase di validazione su dati non precedentemente visti, affetti da rumore gaussiano ($\sigma = 0.1$), al fine di valutare la robustezza predittiva degli algoritmi in condizioni di lieve perturbazione.

La Figura 4.39 presenta un confronto diretto tra i modelli in termini di accuratezza media, evidenziando la percentuale complessiva di classificazioni corrette rispetto al totale delle osservazioni, mentre la Figura 4.40 riporta i valori di F1-score macro, metrica particolarmente utile per valutare la capacità del modello di bilanciare precisione e recall nelle classificazioni multi-classe. La Figura 4.41, infine, rappresenta la relazione tra precisione e recall macro, e permette di valutare l'equilibrio tra la capacità del modello di non produrre falsi positivi (precisione) e quella di rilevare correttamente le classi positive (recall).

L'analisi congiunta di queste metriche evidenzia una sostanziale coerenza con quanto osservato in fase di addestramento, confermando una generale moderata capacità predittiva degli algoritmi in presenza di lieve perturbazione dei dati. In particolare, Logistic Regression e SVM emergono come i modelli più robusti, mantenendo valori di accuratezza e F1-score leggermente superiori rispetto agli altri algoritmi testati. Tuttavia, i valori complessivi di accuratezza e F1-score si mantengono prevalentemente al di sotto della soglia di 0.6, indicando una certa difficoltà nella distinzione netta delle diverse tipologie di latte UHT sulla base dei dati metabolici analizzati.

Il grafico Precisione vs Recall conferma ulteriormente questa osservazione, mostrando una distribuzione piuttosto ravvicinata dei modelli nell'area centrale dello spazio grafico, senza che nessun algoritmo emerga chiaramente per equilibrio tra precisione e recall. Tale risultato riflette probabilmente la limitata variabilità composizionale intrinseca della matrice latte UHT, caratterizzata da processi produttivi altamente standardizzati, che rende la discriminazione tra le diverse classi più complessa rispetto ad altre matrici agroalimentari più variabili.

In conclusione, questi risultati preliminari, sebbene moderati, rappresentano un punto di partenza utile per successive ottimizzazioni del processo analitico. La valutazione definitiva delle performance predittive e della reale capacità di generalizzazione dei modelli sarà pertanto effettuata su un set di dati esterno non perturbato artificialmente, consentendo una stima più accurata della robustezza e della potenzialità applicativa degli algoritmi in contesti reali.

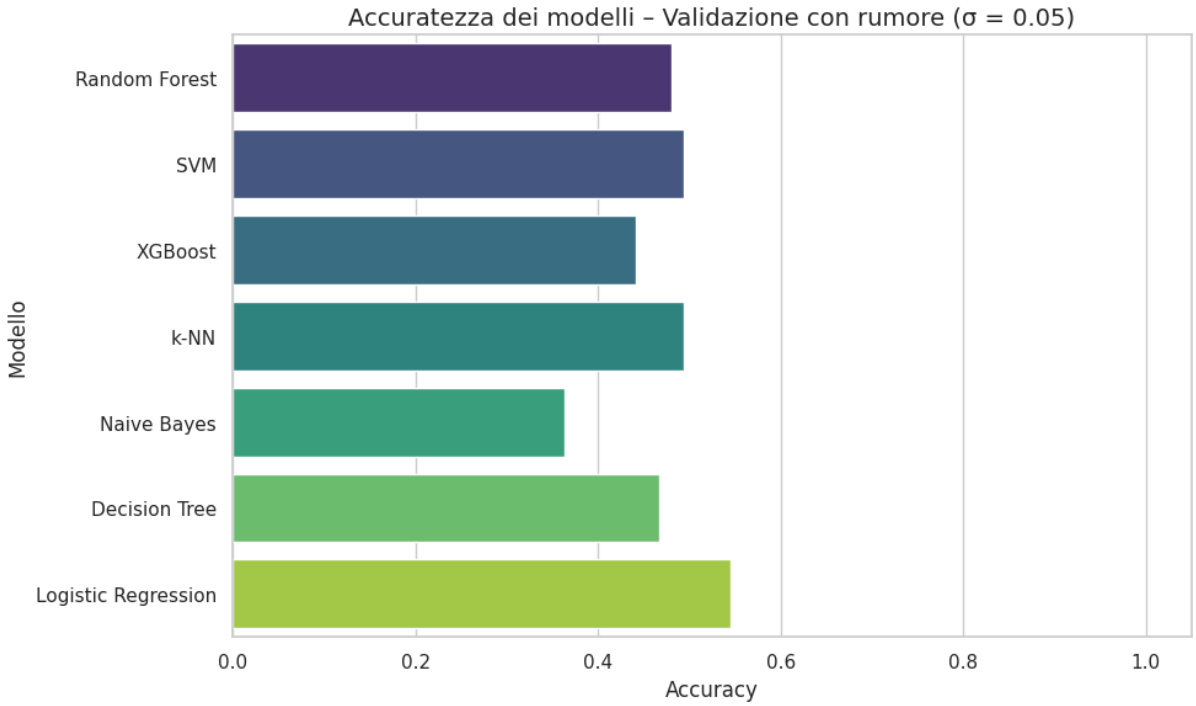


Figura 4.39 – Latte UHT - Accuracy per ciascun modello durante predizione su dati non visti

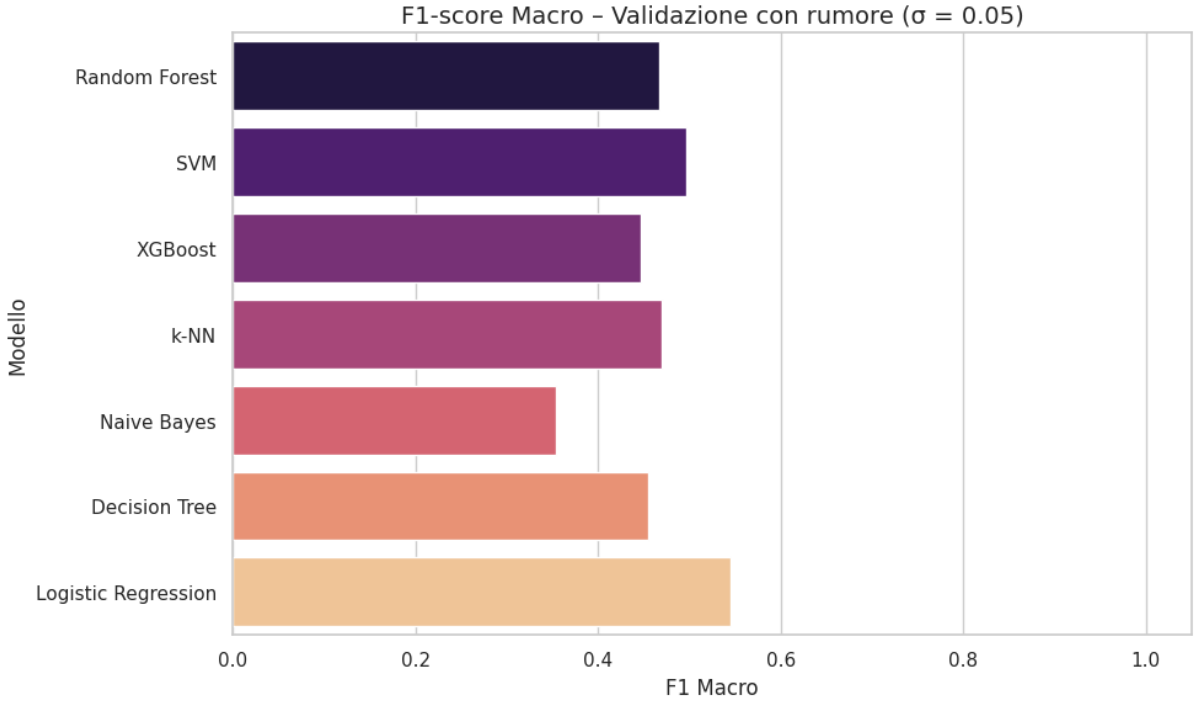


Figura 4.40 – Latte UHT - F1-score Macro per ciascun modello (con rumore gaussiano) durante predizione su dati non visti

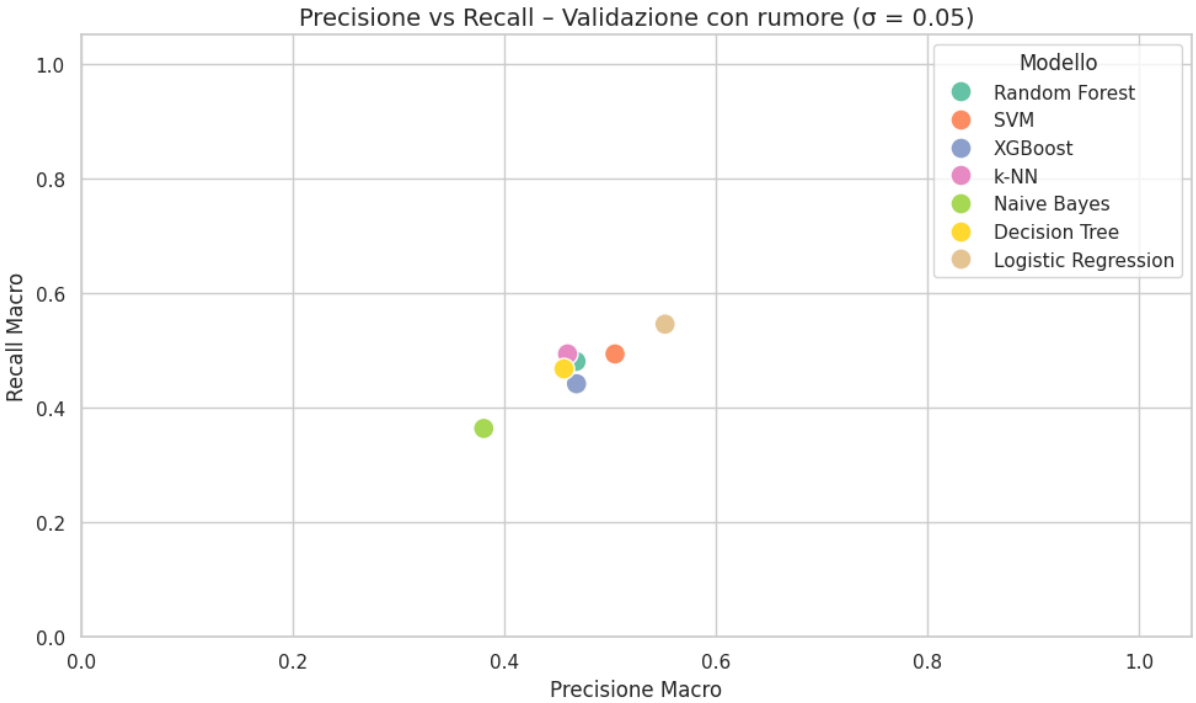


Figura 4.41 – Latte UHT - Precisione vs Recall macro per ciascun modello (con rumore gaussiano) durante predizione su dati non visti

Classificazione finale e selezione delle feature più rilevanti

I risultati riportati nella Tabella 4.19 evidenziano le prime 20 variabili (buckets NMR) che hanno mostrato la maggiore rilevanza durante il processo di classificazione condotto dai diversi algoritmi di machine learning, valutata mediante il criterio della *Mean Decrease Accuracy* (MDA). L'analisi si riferisce alla fase di test dei modelli eseguita su dati affetti da rumore gaussiano di intensità moderata ($\sigma = 0.1$), al fine di simulare condizioni realistiche di variabilità sperimentale.

Le variabili sono elencate in ordine decrescente rispetto al valore massimo di importanza MDA registrato, e sono accompagnate dagli intervalli spettrali corrispondenti (espressi in ppm) e dal numero di modelli predittivi nei quali sono risultate significative per la classificazione.

Tabella 4.19 – Latte UHT - Principali variabili (buckets NMR) selezionate tramite MDA durante il processo di classificazione.

Num	Variabile	PPM		Totale modelli	Importanza MDA
1	Var148	3.50	±0.02	4	1.133
2	Var156	3.18	±0.02	4	1.121
3	Var125	4.42	±0.02	5	1.107
4	Var200	1.42	±0.02	5	1.095
5	Var215	0.82	±0.02	6	1.085
6	Var199	1.46	±0.02	5	1.078
7	Var181	2.18	±0.02	6	1.075
8	Var154	3.26	±0.02	3	1.073
9	Var201	1.38	±0.02	4	1.070
10	Var194	1.66	±0.02	4	1.066
11	Var202	1.34	±0.02	5	1.041
12	Var178	2.30	±0.02	4	1.041
13	Var149	3.46	±0.02	3	1.036
14	Var212	0.94	±0.02	4	1.021
15	Var195	1.62	±0.02	5	1.021
16	Var193	1.70	±0.02	4	1.020
17	Var111	5.18	±0.02	6	1.012
18	Var211	0.98	±0.02	4	1.008
19	Var198	1.50	±0.02	7	1.006
20	Var107	5.34	±0.02	1	1.005

Le prime 20 variabili selezionate mostrano valori di MDA compresi tra 1.005% e 1.133%, delineando un intervallo ristretto che indica un'influenza complessivamente omogenea nel determinare l'accuratezza dei modelli. L'assenza di valori estremamente elevati suggerisce che la classificazione del latte UHT si basi su un insieme distribuito di segnali, senza che emerga una singola variabile dominante. Questo risultato è in linea con la natura della matrice analizzata: il latte UHT rappresenta infatti un prodotto fortemente standardizzato, sia dal punto di vista della composizione che dei trattamenti tecnologici, e pertanto caratterizzato da una limitata variabilità spettrale tra classi.

In questo contesto, il numero di modelli in cui ciascuna variabile è stata selezionata come rilevante diventa un criterio fondamentale per valutare la robustezza delle informazioni spettrali. Alcune variabili, pur con valori di MDA relativamente contenuti, mostrano una notevole stabilità predittiva. Ad esempio, Var198 (1.50 ppm) è stata selezionata da tutti e sette i modelli con un MDA di 1.006%, mentre Var215 (0.82 ppm) e Var181 (2.18 ppm) sono state incluse in sei modelli, con MDA pari rispettivamente a 1.085% e 1.075%. Anche Var111 (5.18 ppm), presente in sei modelli con MDA = 1.012%, rientra tra i segnali più costanti. La selezione

ricorrente di queste variabili, indipendentemente dall'algoritmo utilizzato, suggerisce che esse costituiscano i riferimenti più affidabili nel pattern spettrale di una matrice così omogenea.
Altri segnali, come Var125, Var200, Var199, Var202 e Var195, sono risultati selezionati da cinque modelli, e quindi anch'essi mostrano una buona ricorrenza pur non raggiungendo la trasversalità massima. Diversamente, alcune variabili come Var107 (5.34 ppm), pur presentando un MDA paragonabile (1.005%), sono state individuate da un solo modello, suggerendo una rilevanza più marginale o dipendente da specifici aspetti algoritmici o statistici. Lo stesso vale per variabili come Var154, Var149 e Var178, emerse in tre o quattro modelli, e pertanto meno consistenti nel supportare la classificazione in modo generalizzato.
I risultati ottenuti riflettono pertanto le caratteristiche della matrice latte UHT, ossia una struttura compositiva poco variabile che non favorisce discriminazioni nette tra le classi, ma richiede invece l'identificazione di segnali debolmente differenzianti, il cui valore predittivo si rafforza solo quando considerati nel loro insieme
La combinazione di valori MDA omogenei e selezione ripetuta nei modelli consente comunque di evidenziare alcune variabili robuste e potenzialmente utili per la classificazione.

Validazione mediante set di dati esterno

Nella tabella seguente è riportata la performance dei diversi modelli in termini di accuratezza nella predizione su un set di dati esterno, privo di rumore gaussiano, ottenuto come descritto nel paragrafo 3.2.2.1.
In Figura 4.42 viene presentata, a titolo esemplificativo, la matrice di confusione relativa al modello Random Forest ottenuta nella fase di predizione sul set esterno.

Tabella 4.20 – Latte UHT - Accuratezza, F1-score macro e F1-score weighted dei modelli di classificazione applicati al set esterno di dati.

Modello	Accuracy	F1_macro	F1_weighted
Random Forest	0.961	0.960	0.960
SVM	0.831	0.829	0.829
XGBoost	0.818	0.814	0.814
k-NN	0.727	0.724	0.724
Naive Bayes	0.701	0.695	0.695
Decision Tree	0.909	0.909	0.909
Logistic Regression	0.818	0.817	0.817

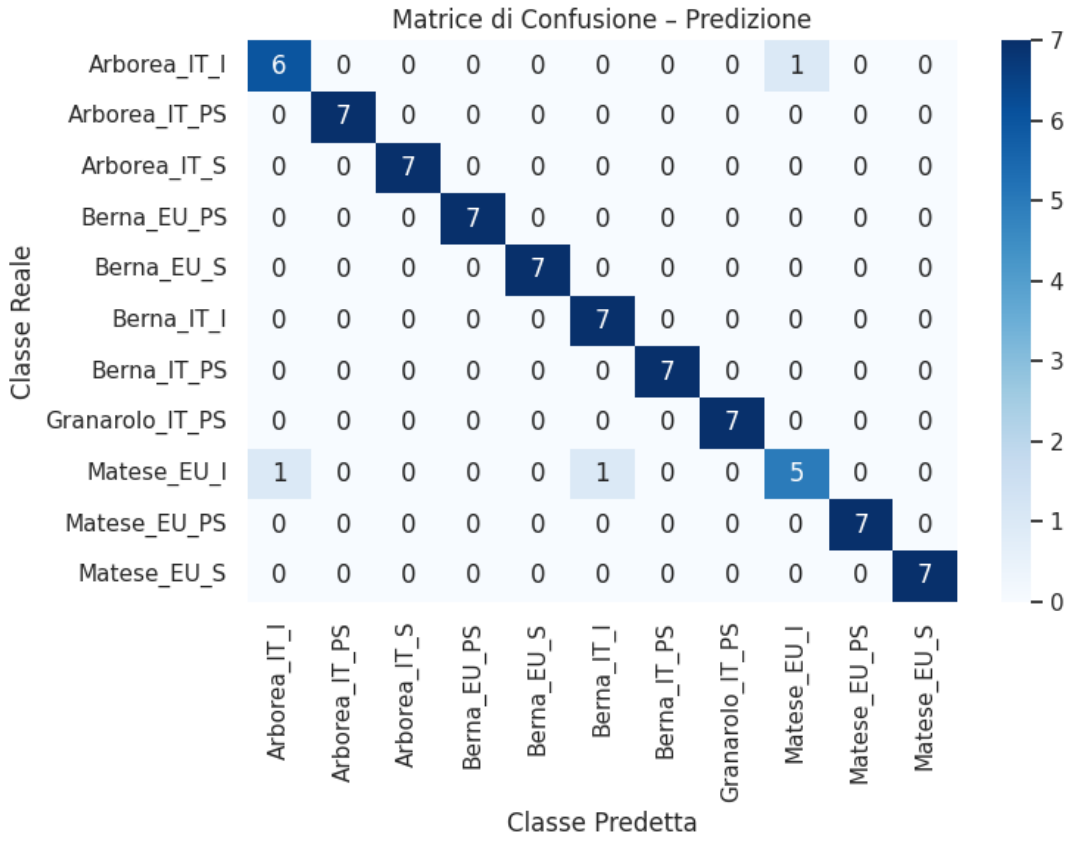


Figura 4.42 – Latte UHT - Matrice di confusione ottenuta con il modello Random Forest durante classificazione con set esterno di dati.

I risultati con predizione su set esterno evidenziano come, tra gli algoritmi considerati, il modello Random Forest abbia ottenuto le performance più elevate, con un'accuratezza del 96,1% e valori quasi identici per F1-score macro e weighted (0.960), a testimonianza di un'elevata capacità di generalizzazione e di una classificazione bilanciata tra le diverse classi. La matrice di confusione associata conferma un'assegnazione quasi perfetta delle osservazioni. Anche il modello Decision Tree ha mostrato buone prestazioni, con accuratezza del 90,9%, segnalando che gli algoritmi ad albero risultano efficaci nel contesto di dati strutturati e ben separabili, come quelli ottenuti dopo riduzione dimensionale. Le performance di SVM e Logistic Regression, sebbene inferiori a quelle dei modelli ad albero, si attestano su valori comunque soddisfacenti (accuratezza ~81–83%), con una buona coerenza tra le metriche F1_macro e F1_weighted.

Prestazioni più contenute sono state invece registrate per XGBoost (81,8%), k-NN (72,7%) e Naive Bayes (70,1%), che risultano meno efficaci nel contesto del set esterno. Questi risultati confermano quanto già osservato durante la fase di validazione interna: l'efficienza predittiva

degli algoritmi tende a rispecchiare la loro capacità di modellare la variabilità intrinseca del dataset, con Random Forest e Decision Tree che si distinguono per robustezza e adattabilità. Nel complesso, la fase di test esterno rafforza la validità del protocollo adottato, confermando la stabilità dei modelli più performanti anche in condizioni realistiche e non simulate.

Discussione finale

L'analisi AI-driven applicata al latte UHT ha evidenziato alcune peculiarità legate alla natura fortemente standardizzata della matrice.

Durante la fase di addestramento, i modelli supervisionati hanno riportato prestazioni complessivamente moderate (accuracy tra 0.41 e 0.60), condizionate sia dall'omogeneità dei dati sia dall'aggiunta di rumore gaussiano. In questo contesto, SVM e Logistic Regression si sono rivelati i più robusti, ma senza superare in modo netto la soglia del 60% di accuratezza. I risultati delle metriche (F1-score e precision-recall) hanno confermato un andamento bilanciato ma contenuto, coerente con la bassa discriminabilità del set.

L'analisi delle feature ha evidenziato un gruppo di variabili con MDA compresi tra 1.005% e 1.133%, distribuite in modo omogeneo e selezionate da più modelli. Variabili come Var198 (1.50 ppm) e Var181 (2.18 ppm), emerse in 6–7 modelli, si confermano tra i segnali più informativi per la classificazione.

La validazione esterna ha invece mostrato un netto miglioramento delle performance: Random Forest ha raggiunto un'accuratezza del 96%, seguito da Decision Tree (91%). Anche SVM e Logistic Regression si sono mantenuti su livelli soddisfacenti (81–83%). Tali risultati indicano che, pur in presenza di prestazioni limitate su dati perturbati, i modelli più performanti riescono a generalizzare efficacemente su campioni reali non manipolati,

4.4.4 Confronto tra approccio classico e automatizzato

Il confronto tra l'approccio chemiometrico convenzionale e quello automatizzato AI-driven, applicati all'analisi del latte UHT, mette in luce vantaggi complementari e limiti specifici legati alla particolare natura della matrice esaminata. L'analisi PCA condotta secondo il metodo classico ha evidenziato una separazione chiara tra le tre tipologie di latte (intero, parzialmente scremato e scremato) lungo la componente F2, associata al contenuto lipidico. Tale risultato è coerente con le osservazioni spettrali preliminari, che mostravano una progressiva attenuazione dei segnali lipidici con la diminuzione della frazione grassa. La selezione delle variabili significative, supportata da ANOVA e test di Tukey, ha permesso l'identificazione di segnali chiave localizzati in regioni ben note, quali i gruppi metilici e metilenici (~0.9–1.4 ppm), i

segnali zuccherini (~3.2–4.5 ppm) e gli amminoacidi aromatici (~6.8–7.4 ppm). L'assegnazione metabolitica ha confermato il ruolo discriminante di composti come fenilalanina, istidina, acido lattico, glucosio, colina e acido butirrico. L'approccio automatizzato, basato su PCA e modelli di machine learning supervisionato, ha mostrato una maggiore difficoltà nel discriminare le tipologie di latte durante la fase di addestramento su dati perturbati: le accuratezze medie sono rimaste generalmente contenute (< 0.60), riflettendo la natura fortemente standardizzata del latte UHT e la bassa variabilità inter-classe. Tuttavia, la successiva classificazione su un set esterno non affetto da rumore ha ribaltato parzialmente questo scenario: modelli come Random Forest (accuracy = 0.961) e Decision Tree (accuracy = 0.909) hanno evidenziato un'elevata capacità di generalizzazione, con performance sensibilmente superiori rispetto alla fase di training. Questo risultato suggerisce che i modelli siano riusciti a cogliere pattern informativi latenti, nonostante la difficoltà iniziale dovuta alla scarsa eterogeneità e alla presenza di rumore artificiale. Dal punto di vista della selezione delle variabili, entrambi gli approcci hanno evidenziato la rilevanza di segnali localizzati nella regione δ 1.2–1.5 ppm (acidi grassi a corta catena, lattato, etanolo), e nella regione δ 3.2–3.5 ppm (glucosio, colina, inositolo), mostrando una convergenza su segnali noti per la loro associazione al contenuto lipidico e al metabolismo energetico. Variabili come Var154 (3.26 ppm), Var202 (1.34 ppm) e Var181 (2.18 ppm) sono state selezionate in entrambi i metodi, suggerendo una coerenza nella rilevazione di marcatori utili, pur in presenza di criteri di selezione diversi. Come nel caso precedente l'approccio classico si è rivelato più efficace nella descrizione biochimica della matrice, mentre il workflow AI-driven ha dimostrato, seppur con iniziali difficoltà, una buona capacità predittiva su dati reali. La convergenza tra i segnali selezionati nei due metodi conferma la solidità delle informazioni spettrali individuate, sebbene l'efficacia discriminativa complessiva risulti più contenuta rispetto ad altre matrici analizzate, a causa della forte omogeneizzazione del latte UHT nei processi industriali. Entrambi gli approcci, dunque, offrono un contributo complementare alla caratterizzazione della matrice, combinando interpretabilità metabolomica e potenzialità predittiva.

4.4.5 Esempio di serializzazione e generazione hash per registrazione del fingerprint metabolico su blockchain

Di seguito viene mostrato un esempio di file serializzato in formato JSON ottenuto a partire dal segnale NMR grezzo (FID) acquisito da un campione rappresentativo della matrice Latte UHT (codice Arborea_IT_I_R1_pD2O), attraverso spettroscopia 1H HR-MAS.


```
{
  "sample_id": "Arborea_IT_I_R1_pD2O_16_05_2022",
  "matrix": "Latte UHT",
  "instrument": "av400",
  "operator": "Cangemi",
  "date": "2022-05-16T15:37:27",
  "temperature": 298.3419,
  "pulse_program": "zgpr",
  "fid_real": [
    0.0,
    0.0,
…
    -856.0
  ],
  "fid_imag": [
    0.0,
    0.0,
...
    -690.0
  ],
  "is_domain": "time",
  "timestamp": "2025-06-10T06:31:59.558299"
}
```


Figura 4.43 – Latte UHT - Esempio di serializzazione da FID Bruker

Il file JSON include la parte reale e quella immaginaria del FID e un insieme di metadati contestuali: strumento, operatore, data, temperatura, sequenza di impulso e timestamp.

Il contenuto del file JSON, sottoposto a funzione di hash SHA256, ha fornito il seguente identificativo digitale:

4e27f82dacd4b3c50aa4aadf826cec902a936d03949d6a03973506a9ccd108d7

Figura 4.44 – Latte UHT - Esempio di generazione hash da FID Bruker

L’hash così ottenuto può fungere da identificativo digitale non modificabile, utile per ancorare il dato analitico a uno smart contract distribuito.
Questo hash rappresenta una "firma digitale" univoca del file, che ne garantisce l’integrità e l’autenticità. Una volta registrato su una rete blockchain (es. Hyperledger Fabric), tale impronta permette di verificare, in qualsiasi momento, che il file non sia stato modificato e che l’acquisizione sia avvenuta in condizioni sperimentali note e tracciabili.

4.5 Cippati

4.5.1 Esempio di integrazione tra spettroscopia NMR e sensoristica IoT per la registrazione in blockchain

4.5.1.1 Acquisizione dei campioni tramite sensori IoT e registrazione dati in blockchain

La fase preliminare di acquisizione dei campioni e registrazione delle relative informazioni su blockchain, mediante l'impiego di sensori IoT, è stata condotta dalla società Farzati S.p.A., nell'ambito del progetto di collaborazione per la presente tesi.

Per ragioni di riservatezza industriale, in questa sezione non si riportano i risultati legati all'acquisizione preliminare e alla generazione del QR code univoco associato a ciascun campione. Si rimanda alla Tabella 3.2 per il solo elenco dei campioni forniti al laboratorio NMR distinti per specie arborea e organizzato secondo la zona geografica e l'altezza di prelievo.

4.5.1.2 Analisi spettroscopica 13C CP-MAS NMR

La Figura 4.45 mostra lo spettro rappresentativo ottenuto da un campione di cippato, nel quale è possibile riconoscere le principali regioni di spostamento chimico (chemical shift), ciascuna corrispondente a specifici gruppi funzionali che riflettono la natura lignocellulosica del materiale.

La regione compresa tra 0 e 45 ppm evidenzia segnali associabili ai carboni alifatici saturi, tra cui spiccano quelli a 20,9 e 32,3 ppm, attribuiti a gruppi metilici e metilenici, presenti soprattutto nei composti lipidici come cere e biooliesteri (cutina e suberina) (Kostryukov et al., 2021). Nella fascia compresa tra 45 e 60 ppm, si osserva un picco marcato a 56,1 ppm, riconducibile ai gruppi metossilici (-OCH_3), componenti strutturali distintivi della lignina, in particolare delle sue subunità guaiaciliche e siringiliche, con possibile sovrapposizione di legami C-N di origine peptidica (Gao et al., 2015).

L'area centrale dello spettro, compresa tra 60 e 110 ppm, è caratterizzata da una forte intensità spettrale, coerente con l'abbondanza di carboidrati strutturali. I picchi spettrali a 64.0 e 72,6 ppm corrispondono rispettivamente ai carboni C6 e alla sovrapposizione dei segnali dei nuclei C2, C3 eC5 dell'anello piranosico del glucosio. Le bande laterali intorno a 85.0 e 89.0 ppm derivano dai C4 impegnati nel legame β 1 $\rightarrow$4 glicosidico con i carboni di-O alchilici anomerici C 1 deschermati a 104,6 ppm. Questa distribuzione conferma la presenza significativa di cellulosa e di emicellulose (Duchesne et al., 2001; Zhu et al., 2016). Tra 110 e 145 ppm la

risonanza a 133.2 ppm identifica i legami C-C e C-H di componenti aromatiche aromatici collegate, alla regione successiva compresa tra 145 e 160 ppm che include i segnali O-aril-C relativi ai carboni fenolici, impegnati nella reticolazione interna della lignina e ai gruppi aromatici del polimero ligninico legati ai sostituenti metossilici e (Conte et al., 2004). Infine, lo spettro si estende fino all'area dei gruppi carbossilici, tra 160 e 220 ppm, dove si osserva un segnale a 171,9 ppm, attribuibile a gruppi carbossili appartenenti a frazioni lipidiche, a componenti della matrice lignocellulosica, quali acidi uronici, e a frazioni ossidate (Kostryukov et al., 2021, 2022).

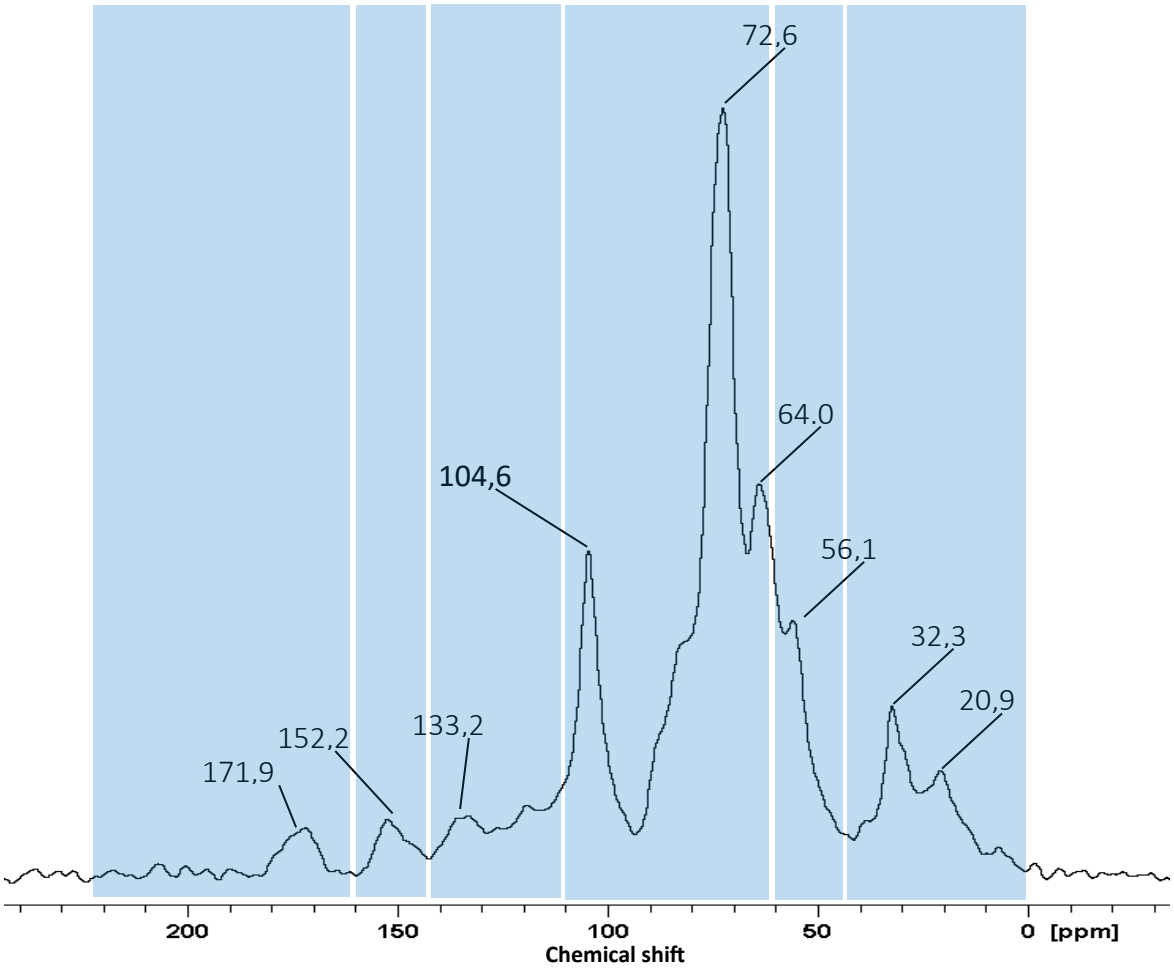


Figura 4.45 – Cippati – Spettro tipico ottenuto da analisi 13C CP-MAS NMR allo stato solido.

Nel complesso, il profilo spettrale fornisce un'indicazione chiara della composizione chimica del cippato, con predominanza della cellulosa e una componente ligninica significativa.
La distribuzione percentuale dei diversi gruppi carboniosi è stata calcolata integrando le aree relative all'intensità dei segnali, in specifici chemical shift, per ogni spettro ^{13}C CP-MAS NMR.

I risultati ottenuti sono riportati nella Tabella 4.21, mentre la loro rappresentazione grafica è mostrata nella Figura 4.46.

Tabella 4.21 - Cippati - Distribuzione percentuale dei gruppi carboniosi nei campioni di legno cippato analizzati mediante spettroscopia 13C CP-MAS NMR. I campioni sono identificati per specie e codice.

	Alkyl-C (0- 45)	Methoxyl-C, C-N (45-60)	O-Alkyl-C (60-110)	Aryl-C (110-145)	Phenol-C (145-160)	Carboxyl-C (160-220)
Abete_Rosso_328	5,2	9,4	70,7	10,9	2,5	1,3
Abete_Rosso_1481	11,9	9,9	62,7	11,7	2,5	1,3
Castagno_173	12,4	8,7	58,9	13,2	4,2	2,7
Castagno_252	12,9	9,2	63,8	9,2	2,9	2,1
Cerro_197	9,0	10,8	64,9	11,0	3,0	1,4
Cerro_243	14,5	10,2	62,1	7,6	3,0	2,7
Cerro_321	9,8	10,3	67,5	8,2	2,6	1,7
Douglasia_238	16,9	11,6	59,0	6,8	2,6	3,1
Douglasia_342	14,7	8,9	59,9	8,8	3,4	4,2
Faggio_141	6,9	8,9	71,4	7,6	2,7	2,6
Faggio_340	12,3	9,4	60,5	11,1	4,5	2,3
Faggio_346	7,3	11,0	67,2	10,1	2,4	2,0
Pino Laricio_217	11,4	10,0	65,9	8,2	2,8	1,7
Pino Laricio_1383	14,1	8,5	57,5	14,1	4,0	1,8
Pino nero_256	10,0	10,2	65,4	9,5	2,8	2,1
Pino_Nero_1348	16,9	9,2	59,7	9,1	2,8	2,3
Pino_Nero_1360	8,7	9,7	69,6	8,0	1,9	2,2
Pioppo_139	13,3	10,9	62,4	8,7	2,3	2,5
Pioppo_146	11,8	9,8	63,0	8,9	2,5	3,9
Pioppo_327	11,8	10,3	64,4	8,5	2,7	2,3

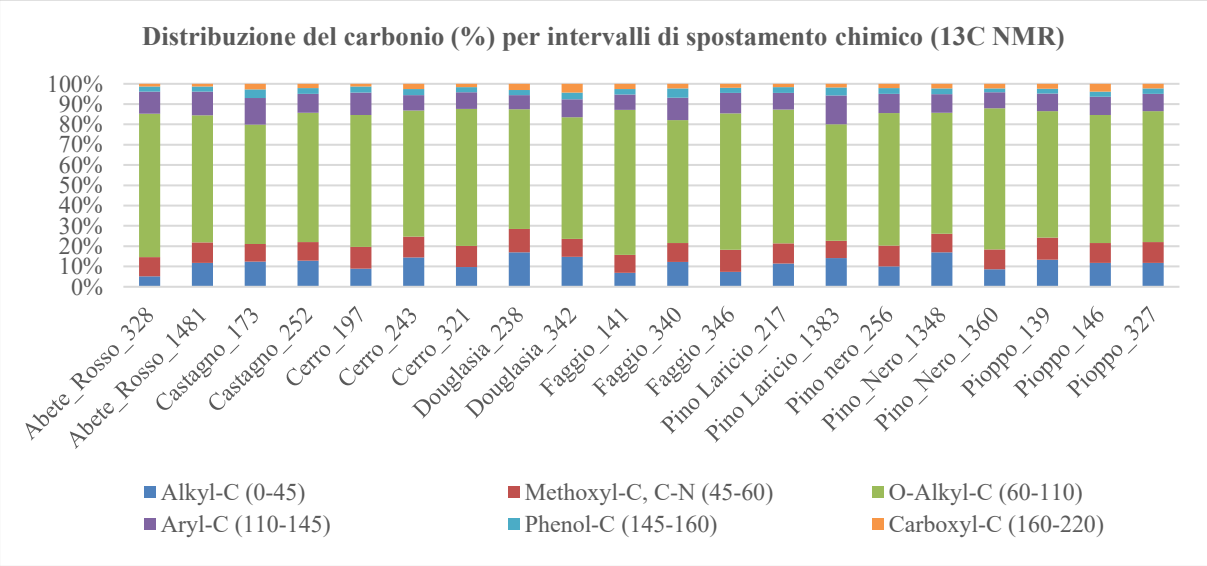


Figura 4.46 – Cippati - Distribuzione percentuale dei gruppi carboniosi nei diversi campioni.

L’analisi dei grafici (Figura 4.47 - Figura 4.49) consente di approfondire il confronto tra i campioni. Nello specifico la regione dei carboni alchilici (Figura 4.47), associata ai carboni alifatici saturi, mostra una marcata variabilità tra le specie. I campioni di Douglasia 238, Pino nero 1348 e Cerro 243 presentano elevati valori (> al 14%), suggerendo una forte predominanza di catene laterali apolari, probabilmente legate a segmenti ligninici meno condensati o a residui lipidici. Al contrario, Abete Rosso 328 e Faggio 346 mostrano segnali riconducibili ad una struttura più ricca in componenti ossigenati. La frazione dei carboni metossilici e carboni legati ad azoto (Figura 4.48) risulta piuttosto uniforme tra i campioni, con valori compresi tra 8% e 12%. Douglasia 238 e Cerro 197 mostrano contenuti metossilici leggermente più elevati rispetto agli altri campioni, probabilmente riconducibili di una maggiore presenza di unità siringiliche. La componente O-Alkyl-C (Figura 4.49), che rappresenta i carboni delle catene glucidiche di cellulosa ed emicellulose, costituisce la frazione dominante in tutti i campioni, con valori superiori al 57% e picchi oltre il 70% per campioni come Faggio 141 e Cerro 321. Tali risultati suggeriscono un’elevata conservazione della componente polisaccaridica, con variazioni imputabili a differenze botaniche piuttosto che a degradazione. La distribuzione della frazione relativa ai carboni aromatici della lignina (Figura 4.50), evidenzia una maggiore incidenza nei campioni Castagno 173 e Pino Laricio 1383, entrambi con valori superiori al 13%. Questo dato evidenzia una componente ligninica più abbondante o più aromaticamente densa. Per quanto riguarda la frazione Phenol-C (Figura 4.51), si osserva una minore variabilità tra i campioni, con valori generalmente compresi tra il 2% e il 4.5%. In questo contesto, nei campioni di Castagno 173, Faggio 340 e Pino nero 256, la componente fenolica risulta lievemente più elevata. Tali dati possono indicare una lignina più reattiva o maggiormente coinvolta in meccanismi di ossidazione e reticolazione. Infine, la frazione Carboxyl-C (Figura 4.52) mostra valori più alti nei campioni Douglasia 342, Pioppo 146 e Castagno 173, suggerendo una maggiore presenza di gruppi ossidati, potenzialmente derivanti da porzioni degradate della lignina o da acidi uronici delle emicellulose. Questo aspetto potrebbe riflettere condizioni ambientali diverse durante la crescita o un grado di maturazione più avanzato. Nel complesso, le differenze evidenziate nei profili spettroscopici non solo riflettono l’eterogeneità botanica dei campioni di cippato, ma suggeriscono anche potenziali correlazioni con la gestione forestale o lo stadio fisiologico del legno.

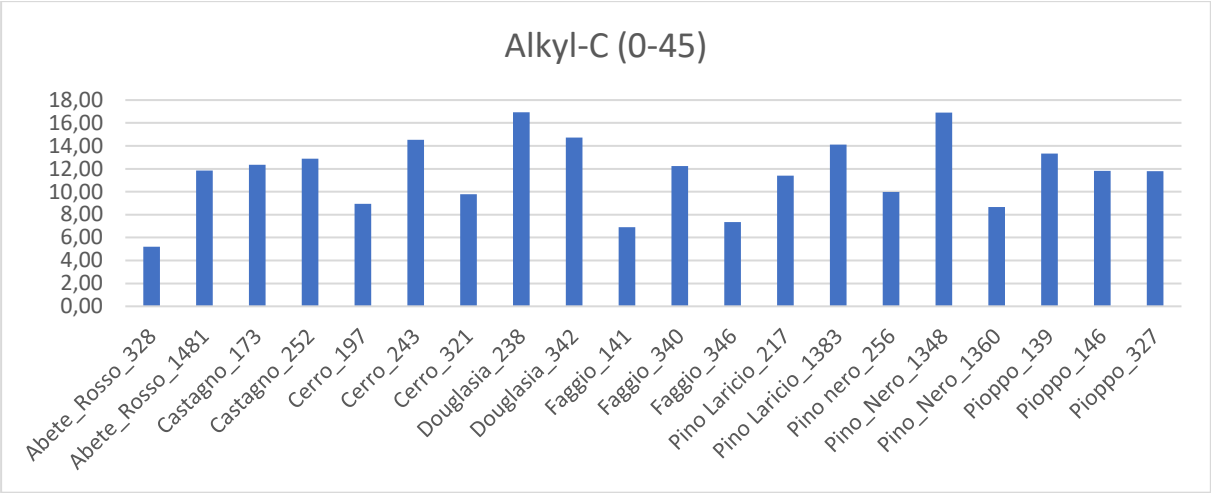


Figura 4.47 – Cippati - Distribuzione percentuale della frazione Alkyl-C (0–45 ppm), rappresentativa dei carboni alifatici saturi.

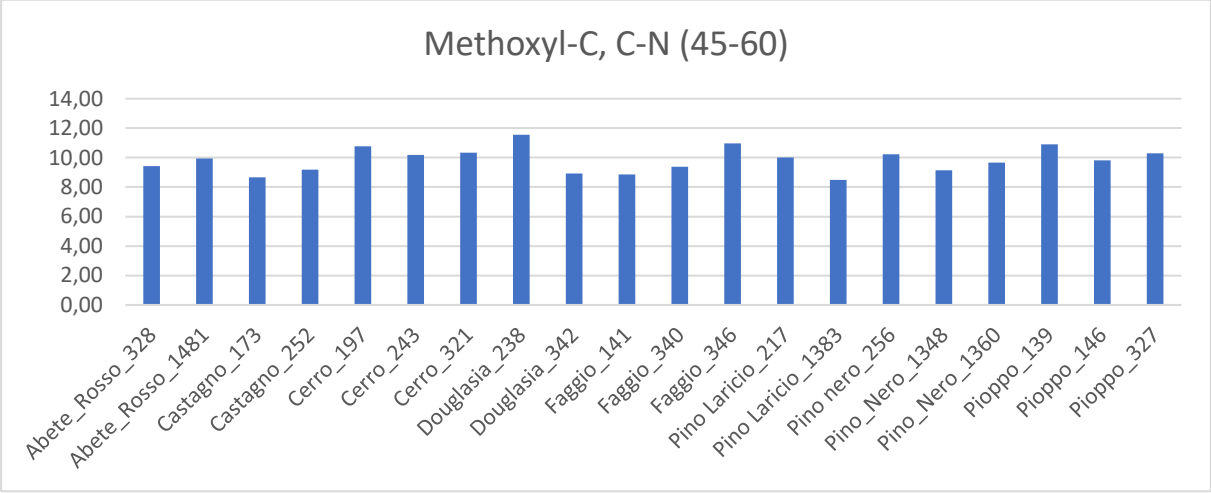


Figura 4.48 – Cippati -Distribuzione percentuale della frazione Methoxyl-C, C-N (45–60 ppm), associata ai gruppi metossilici presenti nella lignina guaiacilica e siringilica.

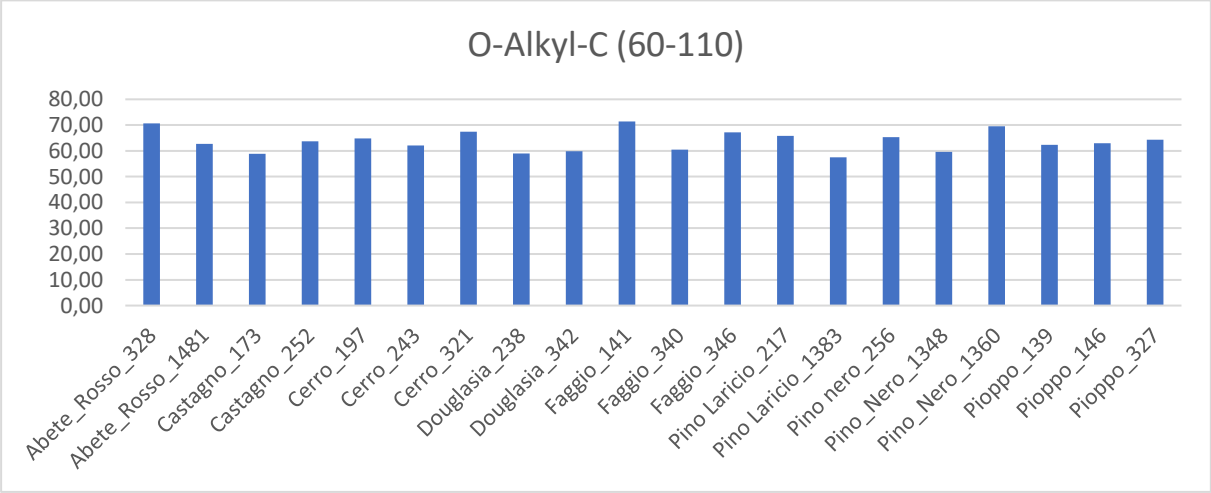


Figura 4.49 – Cippati - Distribuzione percentuale della frazione O-Alkyl-C (60–110 ppm), indicativa della presenza di polisaccaridi strutturali (cellulosa ed emicellulose).

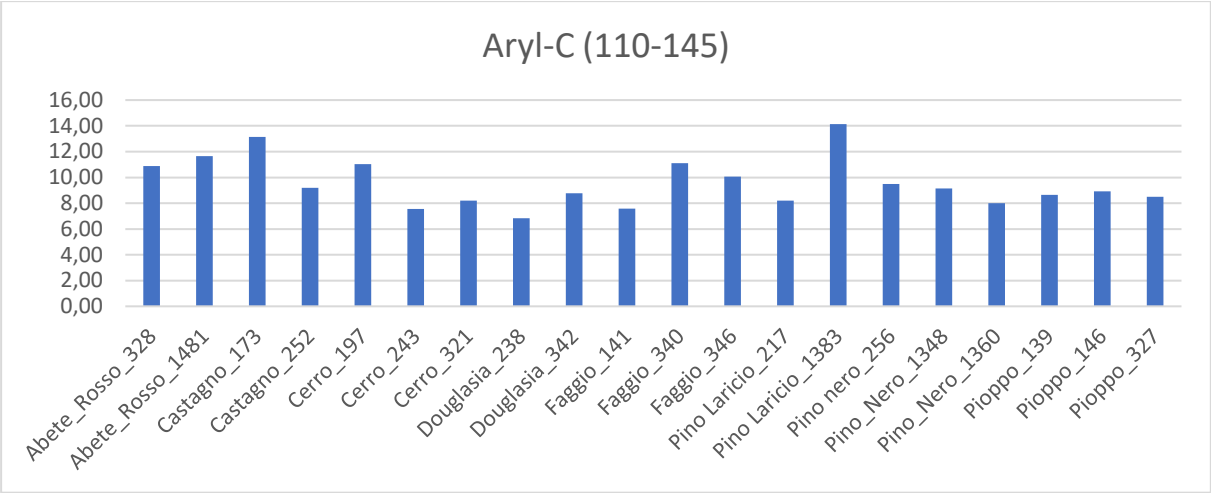


Figura 4.50 – Cippati - Distribuzione percentuale della frazione Aryl-C (110–145 ppm), corrispondente ai carboni aromatici della lignina.

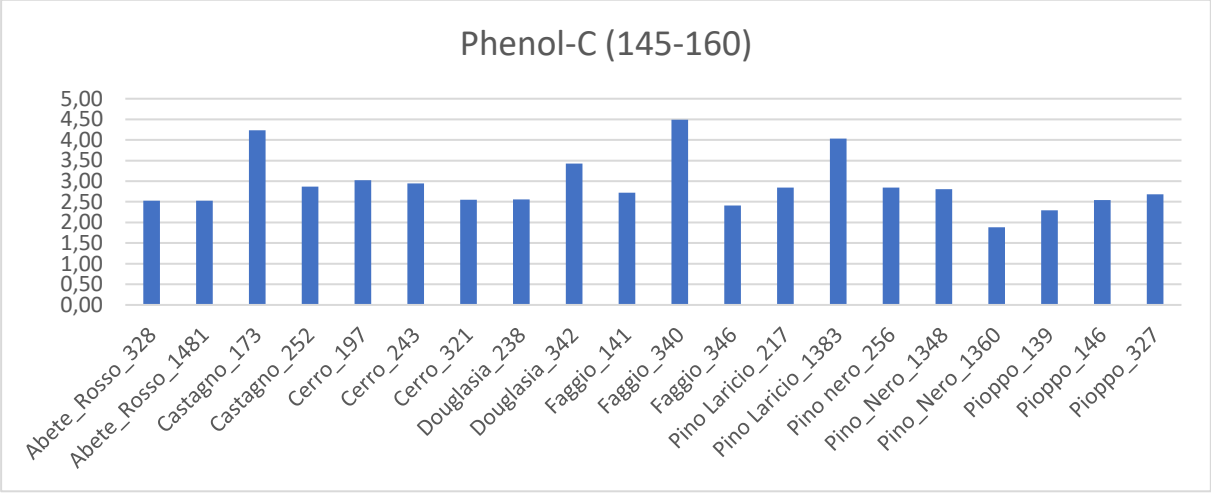


Figura 4.51 – Cippati - Distribuzione percentuale della frazione Phenol-C (145–160 ppm), associata a carboni fenolici sostituiti.

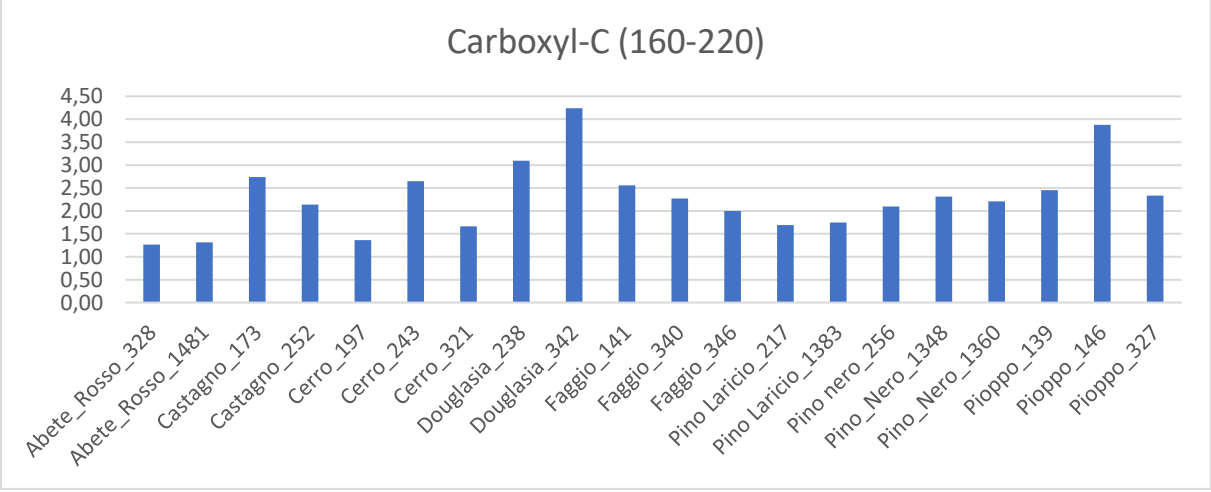


Figura 4.52 – Cippati - Distribuzione percentuale della frazione Carboxyl-C (160–220 ppm), indicativa di componenti ossidati della lignina o acidi uronici da emicellulose.

4.5.1.3 Analisi gerarchica delle similarità tra campioni e aree geografiche

Per approfondire ulteriormente la variabilità compositiva dei cippati, è stata eseguita un'analisi gerarchica basata sulle percentuali relative dei diversi gruppi carboniosi identificati mediante spettroscopia ^{13}C CP-MAS NMR.

La Figura 4.53 mostra il dendrogramma ottenuto dai dati spettrali, con raggruppamenti basati su similarità compositiva tra campioni. Si evidenziano chiaramente tre cluster principali. Il primo gruppo (arancione) include campioni come Faggio 141, Faggio 346, Cerro 197 e Pino nero 256, tutti caratterizzati da una composizione relativamente uniforme con alta presenza di O-Alkyl-C e basse percentuali carboni alchilici. Il secondo cluster (viola) riunisce Castagno 173, Faggio 340, Abete Rosso 1481 e Pino Laricio 1383 suggerendo una maggiore omogeneità nei profili ligninici. Il terzo gruppo (verde) comprende un insieme più eterogeneo, in cui si distinguono campioni con valori elevati di carboni carbossilici o alifatici, come Douglasia 342, Pioppo 146 e Pino nero 1348. Questo raggruppamento più ampio riflette una maggiore varianza intraclasse.

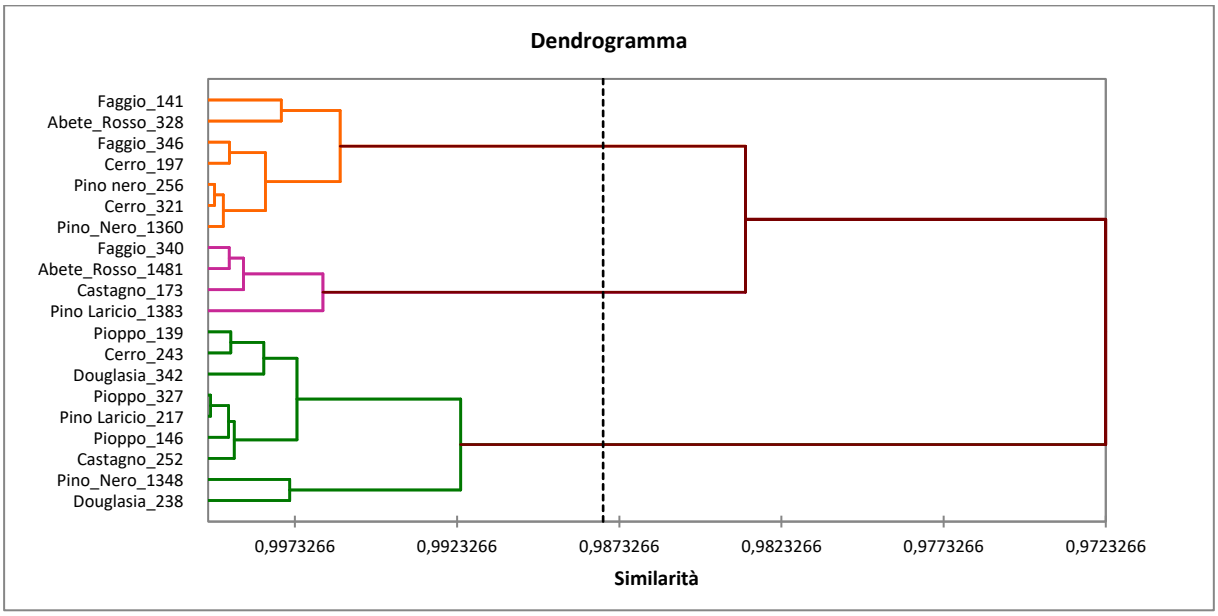


Figura 4.53 – Cippati - Dendrogramma ottenuto dall'analisi gerarchica delle similarità compositive tra campioni di cippato, basata sulla distribuzione percentuale dei gruppi carboniosi (13C CP-MAS NMR).

La Figura 4.54 mostra invece il dendrogramma costruito a partire dai metadati di tracciabilità, in particolare dalle altitudini dei campioni e dalla geolocalizzazione registrata tramite sensore BluDev® e archiviata in blockchain.

Si osserva che sia l'altitudine dei campioni sia la zona di prelievo influenzano la differenziazione dei campioni in gruppi distinti all'interno del dendrogramma, anche quando appartengono alla stessa specie arborea. In particolare, i campioni di Abete rosso si collocano in due cluster differenti, verosimilmente a causa delle diverse caratteristiche chimiche associate alle condizioni ambientali di crescita (altitudini di 1481 m e 328 m) e alle differenti aree di raccolta (Trentino-Alto Adige e Lombardia). Una tendenza analoga è riscontrabile anche per i campioni di Castagno, Pino Laricio e Pino nero, che si distribuiscono in cluster distinti, suggerendo una variabilità molecolare legata all'origine geografica e altitudinale. Al contrario, i campioni di Cerro, Douglasia e Faggio mostrano tra loro alcune differenze chimiche, ma tali variazioni non risultano sufficienti per generare raggruppamenti specifici riconducibili con chiarezza alla zona o all'altitudine di prelievo. I campioni di Pioppo, infine, non evidenziano differenze significative nella composizione chimica. Questi risultati, seppur preliminari, evidenziano l'influenza di fattori ambientali sulla composizione molecolare del cippato e suggeriscono la necessità di ulteriori approfondimenti. Sarà pertanto fondamentale condurre analisi aggiuntive su un numero maggiore di repliche al fine di ottenere evidenze statisticamente solide e validabili.

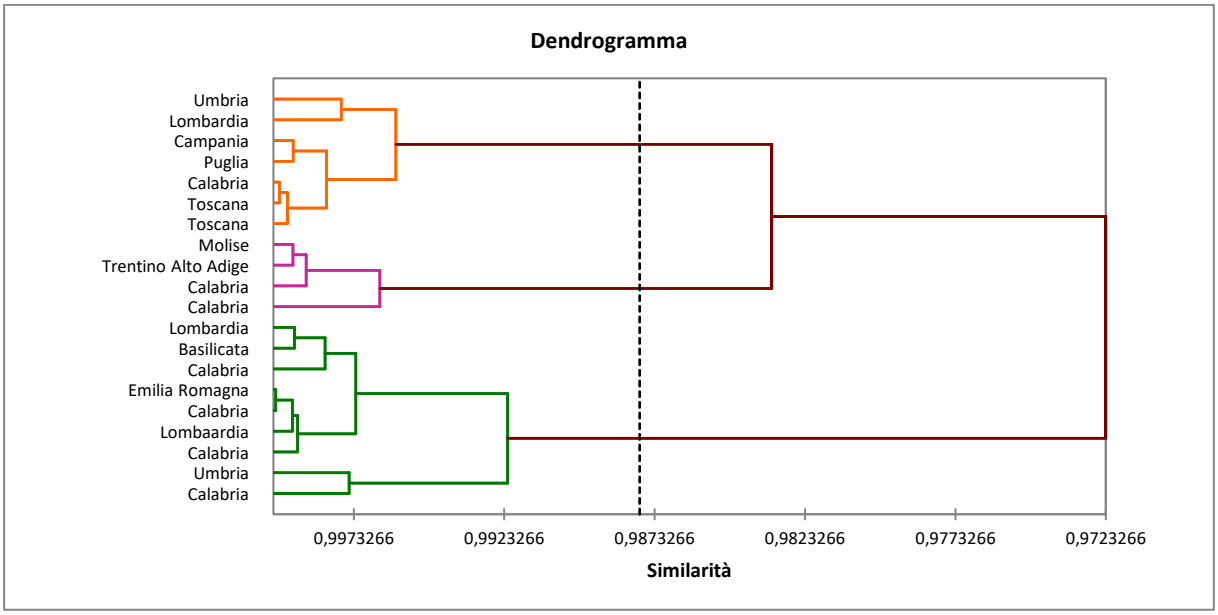


Figura 4.54 – Cippati - Dendrogramma costruito sulla base delle informazioni geografiche di tracciabilità dei campioni, ottenute tramite sensore BluDev® e registrate in blockchain.

Capitolo 5 Conclusioni e Prospettive Future

5.1 Sintesi dei risultati principali

5.1.1 Analisi spettroscopica delle matrici agroalimentari e integrazione con modelli di ML supervisionato

L'impiego della spettroscopia HR-MAS ^{1}H-NMR ha rappresentato la base analitica per l'estrazione di profili metabolici dettagliati da matrici agroalimentari complesse, quali frumento, latte e concentrato di pomodoro. L'elevata risoluzione e riproducibilità della tecnica ha garantito una rappresentazione accurata della composizione chimica dei campioni, fornendo dati idonei a un'elaborazione statistica avanzata. L'integrazione con modelli di machine learning supervisionato ha ampliato significativamente le potenzialità interpretative dell'approccio spettroscopico, permettendo non solo la classificazione efficace dei campioni in base all'origine geografica o alla tipologia, ma anche l'identificazione delle variabili più informative per la discriminazione. L'utilizzo congiunto di metodi di riduzione dimensionale (PCA) e algoritmi di classificazione (Random Forest, SVM, Logistic Regression) ha permesso di costruire modelli predittivi robusti, validati su set esterni e capaci di generalizzare efficacemente anche in condizioni operative realistiche. L'approccio sviluppato si è dimostrato flessibile e replicabile, configurandosi come un paradigma utile per l'analisi integrata di dati spettroscopici e per l'applicazione dell'intelligenza artificiale nel controllo di qualità e nell'autenticazione di prodotti agroalimentari.

5.1.2 Analisi spettroscopica dei Cippati e integrazione con sensoristica IoT per la registrazione in blockchain

L'impiego della spettroscopia 13C CP-MAS NMR per l'analisi dei cippati ha permesso di ottenere un quadro molecolare dettagliato della composizione lignocellulosica dei campioni analizzati, evidenziando le percentuali relative di cellulosa, emicellulose e lignina, nonché le loro variazioni tra specie, zone di provenienza e altitudini. La tecnica NMR allo stato solido, pur presentando alcune limitazioni intrinseche – quali una risoluzione inferiore rispetto alla tecnica HR-MAS e una parziale sovrapposizione dei segnali – si è dimostrata altamente informativa nel rilevare la distribuzione dei gruppi carboniosi e nel descrivere con accuratezza la complessità delle matrici vegetali non trattate.

L'integrazione tra due sistemi consolidati, come la tracciabilità digitale mediante sensore BluDev® (basato su spettroscopia nel vicino infrarosso (NIR), e georeferenziazione GPS e

l'analisi spettroscopica CP-MAS NMR, rappresenta un valore aggiunto strategico in termini di tracciabilità, autenticazione e certificazione del materiale. L'associazione tra il dato chimico ottenuto in laboratorio e quello acquisito direttamente in campo – e immediatamente registrato su rete blockchain – consente di collegare in modo univoco la composizione molecolare del campione alla sua identità geografica, garantendo tracciabilità oggettiva, immutabile e verificabile. Questo approccio integrato, che unisce analisi molecolare ad alta specificità e sistemi avanzati di tracciabilità digitale, offre una soluzione concreta e scalabile per la certificazione di origine, qualità e autenticità dei materiali lignocellulosici, in piena coerenza con i principi della bioeconomia circolare e della gestione sostenibile delle risorse forestali. La possibilità di discriminare – anche solo parzialmente – tra campioni della stessa specie in funzione dell'altitudine o dell'area di raccolta conferma il potenziale applicativo del metodo, sia in ambito produttivo che ambientale. In sintesi, la combinazione tra BluDev® e CP-MAS NMR consente non solo di determinare ciò che il materiale "è" a livello molecolare, ma anche di definirne in modo preciso l'identità spaziale (origine geografica) e temporale (data e ora certificate di raccolta, analisi e registrazione), integrando in un'unica piattaforma il concetto di fingerprinting chimico e tracciabilità geografica certificata.

5.2 Confronto tra approccio classico e AI-driven

Il confronto tra il workflow chemiometrico tradizionale e quello automatizzato AI-driven ha messo in luce differenze sostanziali sia nella strategia analitica sia nelle potenzialità applicative. L'approccio classico, basato sull'analisi PCA, sull'identificazione delle variabili significative tramite ANOVA e test di Tukey e sull'interpretazione spettrale manuale, ha consentito una prima descrizione delle principali fonti di variabilità nei dati NMR e una interpretazione biochimica diretta delle differenze tra classi. Tuttavia, tale metodo presenta limiti in termini di scalabilità, soggettività e capacità predittiva, risultando fortemente dipendente dall'esperienza dell'operatore e dalla struttura del dataset. Al contrario, l'approccio AI-driven ha introdotto un'automazione strutturata del processo analitico, basata su riduzione dimensionale, modellazione supervisionata e selezione delle feature tramite metriche di importanza (MDA). I modelli predittivi sviluppati (Random Forest, SVM, Logistic Regression, ecc.) hanno mostrato, nella maggior parte dei casi, performance elevate anche su dati esterni, con accuratezze spesso superiori al 90% e un'elevata capacità di generalizzazione. Inoltre, la convergenza tra le variabili selezionate nei due approcci ha confermato la robustezza delle informazioni spettroscopiche ottenute: numerosi segnali discriminanti individuati con il metodo classico

sono stati riconosciuti anche dall'analisi automatizzata, rafforzando la validità delle assegnazioni proposte.

5.3 Considerazioni sulla scalabilità e sulle sfide dell'integrazione AI + Blockchain

Sebbene i risultati ottenuti dimostrino la validità dell'approccio AI-driven per la discriminazione varietale e l'identificazione di fingerprint metabolici robusti, l'integrazione effettiva di tali dati in sistemi di certificazione basati su blockchain richiede un'attenta valutazione delle sfide operative, economiche e normative.

Dal punto di vista della scalabilità, l'automazione del workflow analitico (dalla raccolta dello spettro alla classificazione e selezione delle variabili) rappresenta un importante prerequisito per l'adozione su larga scala. Tuttavia, la messa a punto di infrastrutture interoperabili tra laboratori, enti certificatori e nodi blockchain resta un nodo tecnico complesso. La gestione dei costi di analisi, l'aggiornamento dei database metabolici di riferimento e la standardizzazione dei formati di fingerprint da registrare in blockchain sono ulteriori aspetti critici.

Anche il quadro regolatorio e giuridico gioca un ruolo centrale: l'utilizzo di smart contract in ambito agroalimentare implica la definizione di criteri normativi condivisi per la validazione automatica dei dati, il riconoscimento legale delle certificazioni digitali e la protezione dei dati sensibili.

Infine, l'adozione diffusa di queste tecnologie richiede un'elevata fiducia da parte dei soggetti coinvolti (enti di controllo, laboratori, produttori), che può essere ottenuta solo attraverso la trasparenza dei metodi, la riproducibilità dei risultati e la cooperazione tra attori pubblici e privati della filiera.

In quest'ottica, i risultati presentati in questa tesi rappresentano una prova di concetto concreta, utile per avviare progetti pilota e testare l'efficacia dell'integrazione tra AI e blockchain in scenari applicativi reali, con l'obiettivo di contribuire alla costruzione di sistemi alimentari più trasparenti, sicuri e sostenibili.

5.4 Analisi comparativa dei casi studio: sintesi metodologica, potenzialità e limiti applicativi

L'approccio automatizzato descritto è stato applicato a quattro distinte matrici agroalimentari - Triticum durum e Triticum monococcum, frumento Senatore Cappelli, concentrato di pomodoro e latte UHT - al fine di valutarne le prestazioni predittive, la flessibilità metodologica

e la possibile applicabilità in contesti di certificazione e tracciabilità. L'analisi comparativa condotta tra i casi studio ha permesso di trarre considerazioni di carattere generale su più dimensioni operative.

Metodologia impiegata

Il protocollo si è rivelato solido e adattabile a sistemi complessi, caratterizzati da elevata variabilità compositiva e disomogeneità spettrale. L'integrazione sequenziale di fasi di preprocessing standardizzato, riduzione dimensionale tramite PCA, addestramento supervisionato con validazione incrociata stratificata, introduzione controllata di rumore (noise injection) e selezione delle feature basata su MDA ha garantito l'affidabilità statistica dei modelli generati, contenendo al tempo stesso il rischio di overfitting. La pipeline si è inoltre dimostrata replicabile e scalabile, a condizione di mantenere coerenza nelle condizioni sperimentali e nel trattamento dei dati.

Potenziale dei protocolli

Le migliori prestazioni predittive sono state osservate nei casi relativi alle varietà di frumento e al concentrato di pomodoro, per i quali l'accuratezza raggiunta su set esterni si è attestata tra il 95% e il 100%, con valori di F1-score macro e weighted elevati e sostanzialmente sovrapponibili. Anche nel caso del latte UHT — matrice notoriamente complessa e affetta da una maggiore eterogeneità intra-classe — i modelli più robusti (in particolare Random Forest e Decision Tree) hanno dimostrato una soddisfacente capacità di generalizzazione, confermando il potenziale del workflow nel trattare contesti critici. Complessivamente, tali risultati avvalorano l'ipotesi che i profili NMR, opportunamente processati, possano fungere da fingerprint metabolici affidabili per finalità di classificazione e autenticazione.

Criticità nella gestione dei dati

Le principali difficoltà sono state riscontrate in dataset caratterizzati da sbilanciamento tra classi o da significativa varianza intraclasse, come nel caso del latte UHT. Tali condizioni hanno richiesto un'attenta gestione dei dati, sia in fase di preprocessing sia nella scelta delle tecniche di validazione. In alcuni casi, la limitata numerosità campionaria per specifiche classi (ad esempio in alcune categorie di pomodoro) ha rappresentato un vincolo all'ottimizzazione dei modelli. Tuttavia, la pipeline ha mostrato una buona resilienza anche in presenza di dati subottimali, grazie all'adozione di strategie adattive e a una selezione robusta delle variabili discriminanti.

Adeguatezza dei dataset rispetto al flusso di lavoro

I risultati ottenuti confermano che l'efficacia del workflow AI-driven è fortemente influenzata dalla qualità dei dati in input. Fattori quali la standardizzazione delle condizioni sperimentali, il bilanciamento tra le classi, la rappresentatività chimico-metabolica e la gestione del rumore costituiscono elementi determinanti per l'affidabilità delle predizioni. I casi in cui tali requisiti sono stati pienamente soddisfatti hanno mostrato risultati eccellenti, mentre situazioni meno controllate hanno comunque restituito modelli interpretabili, anche se con performance più eterogenee.

Compatibilità con sistemi di certificazione basati su blockchain

I fingerprint metabolici ottenuti con la spettroscopia HR-MAS NMR sono intrinsecamente digitalizzabili e idonei alla codifica tramite funzioni hash, per la successiva archiviazione in registri distribuiti (blockchain). L'elevato grado di accuratezza, riproducibilità e coerenza inter-matrice osservato suggerisce che tali profili possano rappresentare un riferimento affidabile per l'identificazione di lotti e prodotti agroalimentari. In un'ottica di implementazione concreta, l'integrazione con sistemi di tracciabilità blockchain richiederà tuttavia un'infrastruttura interoperabile, una standardizzazione a livello di protocolli e costi di implementazione compatibili con le dinamiche delle filiere. Tali considerazioni aprono a scenari promettenti ma richiedono ulteriori approfondimenti in chiave tecnologica e regolatoria.

5.5 Prospettive future

I risultati ottenuti nel presente lavoro offrono una base metodologica solida per l'evoluzione di strategie di tracciabilità molecolare in ambito agroalimentare e ambientale. L'integrazione tra spettroscopia NMR, modelli predittivi di machine learning e tecnologie di registrazione sicura dei dati, come la blockchain, delinea infatti uno scenario promettente per il futuro della certificazione di qualità e origine. Un primo ambito di sviluppo riguarda l'estensione dell'approccio a un numero più ampio di matrici alimentari e vegetali. L'applicazione del protocollo automatizzato a prodotti diversificati – come cereali minori, derivati fermentati del latte, ortofrutta, legumi, bevande fermentate e materiali vegetali grezzi – permetterebbe di validarne ulteriormente la versatilità, mantenendo l'efficacia nella separazione tra classi, nella selezione di biomarcatori e nella predizione accurata dell'origine o del trattamento subito. In tal senso, la pipeline si configura come uno strumento adattabile, capace di affrontare la

crescente esigenza di tracciabilità molecolare lungo tutta la filiera produttiva, dal campo alla tavola. Un secondo asse di potenziamento riguarda il miglioramento della fase interpretativa, attraverso l'arricchimento della libreria di riferimento dei metaboliti. L'impiego congiunto di dati 2D-NMR (es. COSY, TOCSY, HSQC, HMBC) e MS (spettrometria di massa) consentirà una più precisa assegnazione dei segnali, riducendo l'ambiguità derivante dalla sovrapposizione spettrale e permettendo una caratterizzazione più affidabile dei marcatori metabolici. Questo passaggio sarà particolarmente rilevante per confermare la specificità dei segnali identificati mediante algoritmi AI-driven, rafforzando la validità biochimica delle classificazioni ottenute. Parallelamente, la verifica della robustezza temporale del protocollo mediante test su campioni raccolti in più annate, con differenti condizioni agronomiche, climatiche o pedologiche, rappresenterà un passo essenziale verso l'adozione industriale del metodo. Solo attraverso una validazione su scala temporale sarà possibile dimostrare la stabilità dei pattern metabolici selezionati e la capacità dei modelli di generalizzare al di fuori del set sperimentale originario. In una prospettiva più ampia, l'integrazione con altre tecnologie analitiche – come la spettroscopia FTIR, la cromatografia accoppiata a spettrometria di massa (LC-MS), o l'analisi isotopica (IRMS) – potrà consentire l'adozione di approcci multi-omics o di data fusion. Questi strumenti renderanno possibile l'acquisizione di profili molecolari ancora più dettagliati, ampliando la quantità e la qualità delle informazioni disponibili per la costruzione di modelli di classificazione o regressione, nonché per l'interpretazione biochimico-funzionale delle variabili selezionate. Un ulteriore passo avanti è rappresentato dall'applicazione del protocollo in contesti reali, quali consorzi di tutela, aziende produttrici, enti di certificazione e amministrazioni pubbliche. L'adozione della pipeline come sistema di supporto alla certificazione di denominazioni di origine, marchi di qualità o standard di produzione sostenibile potrà contribuire ad accrescere la trasparenza e l'affidabilità delle filiere, soprattutto in scenari ad alto rischio di frode o di contaminazione da prodotti non conformi. In tale ottica si inserisce il contributo innovativo offerto dal caso studio sui cippati lignocellulosici, analizzati mediante spettroscopia ^{13}C CP-MAS NMR in combinazione con sensoristica NIR (BluDev®) e georeferenziazione GPS. L'integrazione tra analisi molecolare e registrazione in blockchain rappresenta un modello avanzato di tracciabilità, in grado di associare in modo univoco un profilo chimico a un'identità geografica verificabile. L'adozione di questo approccio consente non solo la certificazione dell'origine e della qualità dei materiali forestali, ma anche il monitoraggio della sostenibilità e della coerenza dei processi di raccolta e trasformazione, in linea con i principi della bioeconomia circolare e della gestione forestale responsabile. In sintesi, i risultati ottenuti non rappresentano solo un avanzamento tecnico-metodologico, ma

delineano una direzione applicativa concreta per l'evoluzione delle pratiche di tracciabilità agroalimentare e ambientale. La combinazione tra spettroscopia NMR, modellazione predittiva supervisionata e sistemi di registrazione decentralizzata dei dati consente di costruire un protocollo operativo oggettivo, scalabile e scientificamente fondato. La possibilità di associare un'identità molecolare verificabile a ciascun campione – che sia un alimento trasformato, un cereale, un prodotto lattiero-caseario o un materiale lignocellulosico – apre a scenari in cui l'origine, la qualità e la storia produttiva di una materia prima non siano più dichiarazioni basate su autodichiarazioni o documentazione cartacea, ma informazioni strutturate, tracciabili e non alterabili, accessibili lungo tutta la filiera. In quest'ottica, l'integrazione tra analisi spettrale ad alta risoluzione e sensoristica IoT registrata su blockchain non costituisce solo un'evoluzione tecnica, ma una trasformazione epistemologica del concetto stesso di certificazione.

Capitolo 6 Bibliografia

Abreu, A. C., & Fernández, I. (2020). NMR Metabolomics Applied on the Discrimination of Variables Influencing Tomato (Solanum lycopersicum). In Molecules (Vol. 25, Issue 16). MDPI AG. https://doi.org/10.3390/molecules25163738

Adinolfi, F., Vecchio, Y., Masi, M., Mastandrea, G., Lambertini, G., & De Castro, P. (2024).The reform of EU geographical indications: A look at the newly approved Regulation. AIMS Agriculture and Food, 9(2), 693–698. https://doi.org/10.3934/agrfood.2024037

Aleixandre-Tudo, J. L., Castello-Cogollos, L., Aleixandre, J. L., & Aleixandre-Benavent, R. (2022). Chemometrics in food science and technology: A bibliometric study. Chemometrics and Intelligent Laboratory Systems, 222, 104514. https://doi.org/10.1016/j.chemolab.2022.104514

Alhaj Hamoud, Y., AlGarawi, A. M., Okla, M. K., Sheteiwy, M. S., Khalaf, M. H., Alaraidh, I. A., El-Keblawy, A., Abouleish, M., Sandaña, P., Elsadek, E. A., & Shaghaleh, H. (2025). Metabolomic responses of wheat grains to olive mill wastewater and drought stress treatments. Scientific Reports, 15(1), 13963. https://doi.org/10.1038/s41598-025-98547-2

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Andre, C. M., & Soukoulis, C. (2020). Food quality assessed by chemometrics. Foods, 9(7), 945. https://doi.org/10.3390/foods9070945

Antonicelli, M., Triggiani, M., Musio, B., Latronico, M., Mastrorilli, P., & Gallo, V. (2021, March 23). Agroalimentare e intelligenza artificiale: la nuova frontiera nelle certificazioni e nella lotta alle frodi alimentari. Teckneco. https://www.tekneco.it/ambiente/agroalimentare-e-intelligenza-artificiale-la-nuova-frontiera-nelle-certificazioni-e-nella-lotta-alle-frodi-alimentari

Antonucci, F., Figorilli, S., Costa, C., Pallottino, F., Raso, L., & Menesatti, P. (2019). A review on blockchain applications in the agri-food sector. Journal of the Science of Food and Agriculture, 99(14), 6129–6138. https://doi.org/10.1002/jsfa.9912

Atzei, N., Bartoletti, M., & Cimoli, T. (2017). A survey of attacks on Ethereum smart contracts (SoK). In Proceedings of the 6th International Conference on Principles of Security and Trust (POST) (pp. 164–186). Springer. https://doi.org/10.1007/978-3-662-54455-6_8

Augustijn, D., De Groot, H. J. M., & Alia, A. (2021). HR-MAS NMR applications in plant metabolomics. In Molecules (Vol. 26, Issue 4). MDPI AG. https://doi.org/10.3390/molecules26040931

Baker, J. M., Hawkins, N. D., Ward, J. L., Lovegrove, A., Napier, J. A., Shewry, P. R., & Beale, M. H. (2006). A metabolomic study of substantial equivalence of field-grown

genetically modified wheat. Plant Biotechnology Journal, 4(4), 381–392. https://doi.org/10.1111/j.1467-7652.2006.00197.x

Bambina, P., Spinella, A., Lo Papa, G., Chillura Martino, D. F., Lo Meo, P., Cinquanta, L., & Conte, P. (2024). 1H-NMR spectroscopy coupled with chemometrics to classify wines according to different grape varieties and different terroirs. Agriculture, 14(5), 749. https://doi.org/10.3390/agriculture14050749

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics, 17(3), 166–173. https://doi.org/10.1002/cem.785

Bartel, J., Krumsiek, J., & Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. Computational and Structural Biotechnology Journal, 4(5), e201301009. https://doi.org/10.5936/csbj.201301009

Bénard, C., Bourgaud, F., Gautier, H., Grasselly, D., Navez, B., Caris-Veyrat, C., Weiss, M., & Génard, M. (2015). *Metabolomic profiling in tomato reveals diel compositional changes in fruit affected by source–sink relationships . Journal of Experimental Botany, 66*(11), 3391–3401. https://doi.org/10.1093/jxb/erv151

Benevenuto, R. F., Venter, H. J., Zanatta, C. B., Nodari, R. O., & Agapito-Tenfen, S. Z. (2022). Alterations in genetically modified crops assessed by omics studies: Systematic review and meta-analysis. Trends in Food Science & Technology, 120, 325–337. https://doi.org/10.1016/j.tifs.2022.01.002

Berrar, D. (2019). Logistic Regression. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), Encyclopedia of Systems Biology (pp. 1125–1129). Springer. https://doi.org/10.1007/978-1-4419-9863-7_656

Bertelli, D., Marchetti, L., Pellati, F., & Benvenuti, S. (2019). Use of 1H NMR to Detect the Percentage of Pure Fruit Juices in Blends. Molecules, 24(14), 2592. https://doi.org/10.3390/molecules24142592

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful?. In International Conference on Database Theory (pp. 217–235). Springer. https://doi.org/10.1007/3-540-49257-7_15

Biénabe, E., & Marie-Vivien, D. (2017).Institutionalizing geographical indications in southern countries: Lessons from Basmati and Rooibos. World Development, 98, 58–67. https://doi.org/10.1016/j.worlddev.2017.04.004

Bifarin, O. O. (2023). Interpretable machine learning with treebased shapley additive explanations: Application to metabolomics datasets for binary classification. PLoS ONE, 18(5 May). https://doi.org/10.1371/journal.pone.0284315

Bitquery. (2024). Blockchain Data Warehousing: Challenges in Storing Blockchain Data. Recuperato da https://bitquery.io/blog/blockchain-data-warehousing

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. Computers & Operations Research, 135, 105456. https://doi.org/10.1016/j.cor.2021.105456

Bonanomi, G., Chiurazzi, M., Caporaso, S., Del Sorbo, G., Moschetti, G., & Felice, S. (2008). Soil solarization with biodegradable materials and its impact on soil microbial communities. Soil Biology and Biochemistry, 40(7), 1989–1998. https://doi.org/10.1016/j.soilbio.2008.02.009

Botero-Valencia, J., García-Pineda, V., Valencia-Arias, A., Valencia, J., Reyes-Vera, E., Mejia-Herrera, M., & Hernández-García, R. (2025). Machine Learning in Sustainable Agriculture: Systematic Review and Research Perspectives. Agriculture, 15(4), 377. https://doi.org/10.3390/agriculture15040377

Boudaoud, S., Sicard, D., Suc, L., Conéjéro, G., Segond, D., & Aouf, C. (2021). Ferulic acid content variation from wheat to bread. Food Science and Nutrition, 9(5), 2446–2457. https://doi.org/10.1002/fsn3.2171

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics, 2(4), 171–196. https://doi.org/10.1007/s11306-006-0037-z

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the 20th International Conference on Pattern Recognition. https://doi.org/10.1109/ICPR.2010.764

Bumblauskas, D., Mann, A., Dugan, B., & Rittmer, J. (2020). A blockchain use case in food distribution: Do you know where your food has been? International Journal of Information Management, 52, 102008. https://doi.org/10.1016/j.ijinfomgt.2019.09.004

Byeon, Y. S., Lee, D., Hong, Y. S., Lim, S. T., Kim, S. S., & Kwak, H. S. (2020). Comparison of physicochemical properties and metabolite profiling using 1h NMR spectroscopy of korean wheat malt. Foods, 9(10). https://doi.org/10.3390/foods9101436

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics, 20(8-10), 341–351. https://doi.org/10.1002/cem.1006

Calò, F., Girelli, C. R., & Fanizzi, F. P. (2021). Nuclear magnetic resonance spectroscopy in extra virgin olive oil authentication. Comprehensive Reviews in Food Science and Food Safety, 20(4), 3164–3195. https://doi.org/10.1111/1541-4337.13005ift.onlinelibrary.wiley.com

Carmelo Corsaro, Nicola Cicero, Domenico Mallamace, Sebastiano Vasi, Clara Naccari, Andrea Salvo, Salvatore Vincenzo Giofrè, Giacomo Dugo, HR-MAS and NMR towards Foodomics, Food Research International, Volume 89, Part 3, 2016, Pages 1085-1094, ISSN 0963-9969, https://doi.org/10.1016/j.foodres.2016.09.033.

Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. Telematics and Informatics, 36, 55–81. https://doi.org/10.1016/j.tele.2018.11.006

Cassago, A. L. L., Artêncio, M. M., de Moura Engracia Giraldi, J., & Da Costa, F. B. (2021). Metabolomics as a marketing tool for geographical indication products: A literature review. European Food Research and Technology, 247(9), 2143–2159. https://doi.org/10.1007/s00217-021-03782-2

Cei, L., Defrancesco, E., & Stefani, G. (2018).From geographical indications to rural development: A review of the economic effects of European Union policy. Sustainability, 10(10), 3745. https://doi.org/10.3390/su10103745

Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: A review. Trends in Food Science & Technology, 20(11–12), 557–566. https://doi.org/10.1016/j.tifs.2009.07.002

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chi, J., Shu, J., Li, M., Mudappathi, R., Jin, Y., Lewis, F., Boon, A., Qin, X., Liu, L., & Gu, H. (2024). Artificial intelligence in metabolomics: A current review. TrAC - Trends in Analytical Chemistry, 178, Article 117852. https://doi.org/10.1016/j.trac.2024.117852

Chilla, T., Fink, B., Balling, R., Reitmeier, S., & Schober, K. (2020).The EU food label “Protected Geographical Indication”: Economic implications and their spatial dimension. Sustainability, 12(14), 5503. https://doi.org/10.3390/su12145503

Cipriano, D. F., Chinelatto, L. S., Nascimento, S. A., Rezende, C. A., de Menezes, S. M. C., & Freitas, J. C. C. (2020). Potential and limitations of 13C CP/MAS NMR spectroscopy to determine the lignin content of lignocellulosic feedstock. Biomass and Bioenergy, 142, 105792. https://doi.org/10.1016/j.biombioe.2020.105792

Consonni, R., & Cagliani, L. R. (2019). The potentiality of NMR-based metabolomics in food science and food authentication assessment. Magnetic Resonance in Chemistry, 57(9), 558–578. https://doi.org/10.1002/mrc.4807

Consonni, R., Cagliani, L. R., Stocchero, M., & Porretta, S. (2009). Triple concentrated tomato paste: Discrimination between Italian and Chinese products. Journal of Agricultural and Food Chemistry, 57(11), 4506–4513. https://doi.org/10.1021/jf804004z

Conte, P., Spaccini, R., & Piccolo, A. (2004). State of the art of CPMAS 13C-NMR spectroscopy applied to natural organic matter. Progress in Nuclear Magnetic Resonance Spectroscopy, 44(3–4), 215–223. https://doi.org/10.1016/j.pnmrs.2004.02.002

Conti, M. (2022). Blockchain per la tracciabilità dell'olio extravergine d'oliva: soluzioni low-cost per piccoli produttori. Journal of Agricultural Informatics, 13(2), 45–53. https://doi.org/10.1234/jai.2022.002

Corsaro, C., Mallamace, D., Vasi, S., Ferrantelli, V., Dugo, G., & Cicero, N. (2015). ^1H HR-MAS NMR spectroscopy and the metabolite determination of typical foods in Mediterranean diet. Journal of Analytical Methods in Chemistry, 2015, 175696. https://doi.org/10.1155/2015/175696

Corsaro, C., Vasi, S., Neri, F., Mezzasalma, A. M., Neri, G., & Fazio, E. (2022). NMR in Metabolomics: From Conventional Statistics to Machine Learning and Neural Network Approaches. In Applied Sciences (Switzerland) (Vol. 12, Issue 6). MDPI. https://doi.org/10.3390/app12062824

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Analytical Chemistry, 78(7), 2262–2267. https://doi.org/10.1021/ac0519312

Cubero-Leon, E., Peñalver, R., & Maquet, A. (2014). Review on metabolomics for food authentication. Food Research International, 60, 95–107. https://doi.org/10.1016/j.foodres.2013.11.041

Cui, C., Xu, Y., Jin, G., Zong, J., Peng, C., Cai, H., & Hou, R. (2023). Machine learning applications for identify the geographical origin, variety and processing of black tea using ^{1}H NMR chemical fingerprinting. Food Control, 148, 109686. https://doi.org/10.1016/j.foodcont.2023.109686

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Czolpinska, M., & Rurek, M. (2018). Plant glycine-rich proteins in stress response: An emerging, still prospective story. In Frontiers in Plant Science (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fpls.2018.00302

Dasenaki, M. E., Drakopoulou, S. K., Aalizadeh, R., & Thomaidis, N. S. (2019). Targeted and Untargeted Metabolomics as an Enhanced Tool for the Detection of Pomegranate Juice Adulteration. Foods, 8(6), 212. https://doi.org/10.3390/foods8060212

Dupont, F. M. (2008). Metabolic pathways of the wheat (Triticum aestivum) endosperm amyloplast revealed by proteomics. BMC Plant Biology, 8. https://doi.org/10.1186/1471-2229-8-39

Ellis, D. I., Brewster, V. L., Dunn, W. B., Allwood, J. W., Golovanov, A. P., & Goodacre, R. (2012). Fingerprinting food: Current technologies for the detection of food adulteration and contamination. Chemical Society Reviews, 41(17), 5706–5727. https://doi.org/10.1039/c2cs35138b

Eltemur, D., Robatscher, P., Oberhuber, M., Scampicchio, M., & Ceccon, A. (2023). Applications of Solution NMR Spectroscopy in Quality Assessment and Authentication of Bovine Milk. In Foods (Vol. 12, Issue 17). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/foods12173240

Emwas, A. H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., ... & Wishart, D. S. (2019).NMR spectroscopy for metabolomics research. Metabolites, 9(7), 123. https://doi.org/10.3390/metabo9070123

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikström, C. (2013). Multi- and Megavariate Data Analysis – Basic Principles and Applications. Umetrics Academy.

Fordellone, M., Bellincontro, A., & Mencarelli, F. (n.d.). Partial Least Squares Discriminant Analysis: A Dimensionality Reduction Method To Classify Hyperspectral Data.

Francisco, K., & Swanson, D. (2018). The Supply Chain Has No Clothes: Technology Adoption of Blockchain for Supply Chain Transparency. Logistics, 2(1), 2. https://doi.org/10.3390/logistics2010002

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. In Frontiers in Genetics (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fgene.2022.1017340

Gallo, V., Ragone, R., Musio, B. et al. A Contribution to the Harmonization of Non-targeted NMR Methods for Data-Driven Food Authenticity Assessment. Food Anal. Methods 13, 530–541 (2020). https://doi.org/10.1007/s12161-019-01664-8

Galvez, J. F., Mejuto, J. C., & Simal-Gandara, J. (2018). Future challenges on the use of blockchain for food traceability analysis. Trends in Analytical Chemistry, 107, 222–232. https://doi.org/10.1016/j.trac.2018.08.011

García-Pérez, P., Becchi, P. P., Zhang, L., Rocchetti, G., & Lucini, L. (2024). Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity. Trends in Food Science & Technology, 147, Article 104481. https://doi.org/10.1016/j.tifs.2024.104481

George, W., & Al-Ansari, T. (2023). Review of Blockchain Applications in Food Supply Chains. Blockchains, 1(1), 34–57. https://doi.org/10.3390/blockchains1010004

Girelli, C. R., Calò, F., Angilè, F., Mazzi, L., Barbini, D., & Fanizzi, F. P. (2020). ^{1}H NMR spectroscopy to characterize Italian extra virgin olive oil blends, using statistical models and databases based on monocultivar reference oils. Foods, 9(12), 1797. https://doi.org/10.3390/foods9121797

Gowda, G. A. N., & Raftery, D. (2021). NMR-based metabolomics. In Advances in Experimental Medicine and Biology (Vol. 1280, pp. 19–37). Springer. https://doi.org/10.1007/978-3-030-51652-9_2

Guo, F., Lin, G., Dong, L., Cheng, K.-K., Deng, L., Xu, X., Raftery, D., & Dong, J. (2023). Concordance-based batch effect correction for large-scale metabolomics. Analytical Chemistry, 95(1), 123–131. https://doi.org/10.1021/acs.analchem.2c05748

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. In Journal of Machine Learning Research (Vol. 3).

Hanson, A. D., Rathinasabapathi, B., Rivoal, J., Burnet, M., Dillon, M. 0, & Gage, D. A. (1994). Osmoprotective compounds in the Plumbaginaceae: A natural experiment in metabolic engineering of stress tolerance (Vol. 91).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hategan, A. R., Pirnau, A., & Magdas, D. A. (2025). Applications of Machine Learning for Wine Recognition Based on 1H-NMR Spectroscopy. Beverages, 11(2), 45. https://doi.org/10.3390/beverages11020045

Herbke, P., Lamichhane, S., Barman, K., Pandey, S. R., Küpper, A., Abraham, A., & Sabadello, M. (2024). DIDChain: Advancing Supply Chain Data Management with Decentralized Identifiers and Blockchain. arXiv preprint arXiv:2406.11356. https://arxiv.org/abs/2406.11356

Heyman, H. M., & Meyer, J. J. M. (2012). NMR-based metabolomics as a quality control tool for herbal products. South African Journal of Botany, 82, 21–32.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

Hou, L., Zhang, Z., Zhang, Y., Zhang, M., Hu, J., Zheng, Y., & Wang, W. (2024). ^{1}H NMR spectroscopy combined with machine-learning algorithm for origin recognition of Chinese famous green tea Longjing tea. Food Chemistry, 438, 135046. https://doi.org/10.1016/j.foodchem.2024.135046

Hsu, C. W., Chang, C. C., & Lin, C. J. (2010). A practical guide to support vector classification. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Ingallina, C., Sobolev, A. P., Circi, S., Spano, M., Giusti, A. M., & Mannina, L. (2020). New Hybrid Tomato Cultivars: An NMR-Based Chemical Characterization. *Applied Sciences*, *10*(5), 1887. https://doi.org/10.3390/app10051887

Jacobsen, N. E. (2007). NMR Spectroscopy Explained. Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kamath, R. (2018). Food Traceability on Blockchain: Walmart's Pork and Mango Pilots with IBM. The Journal of the British Blockchain Association, 1(1), 1–12. https://doi

Kamilaris, A., Fonts, A., & Prenafeta-Boldú, F. X. (2019). The rise of blockchain technology in agriculture and food supply chains. Trends in Food Science & Technology, 91, 640–652. https://doi.org/10.1016/j.tifs.2019.07.034

Kiani, R., Arzani, A., & Mirmohammady Maibody, S. A. M. (2021). Polyphenols, Flavonoids, and Antioxidant Activity Involved in Salt Tolerance in Wheat, Aegilops cylindrica and Their Amphidiploids. Frontiers in Plant Science, 12. https://doi.org/10.3389/fpls.2021.646221

Kirwan, J. A., Broadhurst, D. I., Davidson, R. L., & Viant, M. R. (2009). Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Analytical and Bioanalytical Chemistry, 394(1), 267–281. https://doi.org/10.1007/s00216-009-2647-7

Kostryukov, S. G., Petrov, P. S., Kalyazin, V. A., Masterova, Y. Y., Tezikova, V. S., Khluchina, N. A., Labzina, L. Y., & Alalvan, D. K. (2021). Determination of lignin content in plant materials using solid-state 13C NMR spectroscopy. Polymer Science, Series B, 63(5), 544–552. https://doi.org/10.1134/S1560090421050067

Koutras, C., Kourkouni, A., & Zampetakis, L. A. (2023). An Ethereum-Based Distributed Application for Enhancing Food Traceability: The Case of Table Olives. Foods, 12(6), 1220. https://doi.org/10.3390/foods12061220

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, S., de Jesus, R. P., & Borges, R. M. (2024). Nuclear Magnetic Resonance and Artificial Intelligence. Encyclopedia, 4(4), 1568–1580. https://doi.org/10.3390/encyclopedia4040102

Lankham, I., & Slaughter, M. (2020). Simple and efficient bootstrap validation of predictive models using SAS/STAT® software. In SAS Global Forum 2020 Proceedings. SAS Institute Inc. https://support.sas.com/resources/papers/proceedings20/4647-2020.pdf

Li, L., Dou, N., Zhang, H., & Wu, C. (2021). The versatile GABA in plants. In Plant Signaling and Behavior (Vol. 16, Issue 3). Bellwether Publishing, Ltd. https://doi.org/10.1080/15592324.2020.1862565

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Liu, C., Chen, F., Liu, L., Fan, X., Liu, H., Zeng, D., & Zhang, X. (2022). The Different Metabolic Responses of Resistant and Susceptible Wheats to Fusarium graminearum Inoculation. Metabolites, 12(8). https://doi.org/10.3390/metabo12080727

Livera, A. M. D., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A., & Speed, T. P. (2015). Statistical methods for handling unwanted variation in metabolomics data. Analytical Chemistry, 87(7), 3606–3615. https://doi.org/10.1021/ac504955r

Ljunggren, D. (n.d.). A Comparative Analysis of Robustness to Noise in Machine Learning Classifiers. In DEGREE PROJECT TECHNOLOGY.

Lukacs, M., Toth, F., Horvath, R., Solymos, G., Alpár, B., Varga, P., Kertesz, I., Gillay, Z., Baranyai, L., Felfoldi, J., Nguyen, Q. D., Kovacs, Z., & Friedrich, L. (2025). Advanced Digital Solutions for Food Traceability: Enhancing Origin, Quality, and Safety Through NIRS, RFID, Blockchain, and IoT. Journal of Sensor and Actuator Networks, 14(1), 21. https://doi.org/10.3390/jsan14010021

Lukman, A. F., Mohammed, S., Olaluwoye, O., & Farghali, R. A. (2025). Handling Multicollinearity and Outliers in Logistic Regression Using the Robust Kibria–Lukman Estimator. Axioms, 14(1), 19. https://doi.org/10.3390/axioms14010019

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Maestrello, V., Solovyev, P., Bontempo, L., Mannina, L., & Camin, F. (2023). Nuclear magnetic resonance spectroscopy in extra virgin olive oil authentication. Comprehensive

Reviews in Food Science and Food Safety, 22(2), 1155–1180. https://doi.org/10.1111/1541-4337.13005

Maestrello, V., Solovyev, P., Franceschi, P., Stroppa, A., & Bontempo, L. (2024). 1H-NMR approach for the discrimination of PDO Grana Padano cheese from non-PDO cheeses. Foods, 13(3), 358. https://doi.org/10.3390/foods13030358

Maestrello, V., Solovyev, P., Stroppa, A., Bontempo, L., & Franceschi, P. (2024). 1H NMR profiling and chemometric analysis for ripening and production characterization of Grana Padano cheese. Food Chemistry, 456, 139986. https://doi.org/10.1016/j.foodchem.2024.139986

Mannina, L., D'Imperio, M., Gobbino, M., D'Amico, I., Casini, A., Emanuele, M. C., & Sobolev, A. P. (2012). Nuclear magnetic resonance study of flavoured olive oils. Flavour and Fragrance Journal, 27(3), 250–259. https://doi.org/10.1002/ffj.2100

Mao, D., Hao, Z., Wang, F., & Li, H. (2018). Innovative blockchain-based approach for sustainable and credible environment in food trade: A case study in Shandong Province, China. Sustainability, 10(9), 3149. https://doi.org/10.3390/su10093149

Märtens, A., Holle, J., Mollenhauer, B., Wegner, A., Kirwan, J., & Hiller, K. (2023). Instrumental Drift in Untargeted Metabolomics: Optimizing Data Quality with Intrastudy QC Samples. Metabolites, 13(5), 665. https://doi.org/10.3390/metabo13050665

Mas, F. D., Massaro, M., Ndou, V., & Raguseo, E. (2023).Blockchain technologies for sustainability in the agrifood sector: A literature review of academic research and business perspectives. Technological Forecasting and Social Change, 187, 122155. https://doi.org/10.1016/j.techfore.2022.122155

Mazzei, P., & Piccolo, A. (2017). HRMAS NMR spectroscopy applications in agriculture. Chemical and Biological Technologies in Agriculture, 4(1), 11. https://doi.org/10.1186/s40538-017-0093-9

Mazzei, P., & Piccolo, A. (2018). NMR-based metabolomics of water-buffalo milk after conventional or biological feeding. Chemical and Biological Technologies in Agriculture, 5(1). https://doi.org/10.1186/s40538-017-0116-6

Mazzei, P., Francesca, N., Moschetti, G., & Piccolo, A. (2010). NMR spectroscopy evaluation of direct relationship between soils and molecular composition of red wines from Aglianico grapes. Analytica Chimica Acta, 673(2), 167–172. https://doi.org/10.1016/j.aca.2010.06.003

Md Nor, S., Ding, P., Abas, F., & Mediani, A. (2022). 1H NMR reveals dynamic changes of primary metabolites in purple passion fruit (Passiflora edulis Sims) juice during maturation and ripening. Agriculture, 12(2), 156. https://doi.org/10.3390/agriculture12020156

Md Nor, S., Ding, P., Abas, F., & Mediani, A. (2022). ^{1}H NMR reveals dynamic changes of primary metabolites in purple passion fruit (Passiflora edulis Sims) juice during maturation and ripening. Agriculture, 12(2), 156. https://doi.org/10.3390/agriculture12020156

Meoni, G., Tenori, L., Di Cesare, F., Brizzolara, S., Tonutti, P., Cherubini, C., Mazzanti, L., & Luchinat, C. (2025). NMR-based metabolomic approach to estimate chemical and sensorial profiles of olive oil. Computational and Structural Biotechnology Journal, 27, 1359–1369. https://doi.org/10.1016/j.csbj.2025.03.045

Mialon, N., Roig, B., Capodanno, E., & Cadiere, A. (2023). Untargeted metabolomic approaches in food authenticity: A review that showcases biomarkers. Food Chemistry, 398, 133856. https://doi.org/10.1016/j.foodchem.2022.133856

Michaletti, A., Naghavi, M. R., Toorchi, M., Zolla, L., & Rinalducci, S. (2018b). Metabolomics and proteomics reveal drought-stress responses of leaf tissues from spring-wheat. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-24012-y

Moco, Sofia. (2022). Studying Metabolism by NMR-Based Metabolomics. Frontiers in Molecular Biosciences. 9. 882487. 10.3389/fmolb.2022.882487.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Montgomery, K. H., Elhabashy, A., Del Carmen Reynoso Rivas, M., et al. (2024). NMR metabolomics as a complementary tool to brix-acid tests for navel orange quality control of long-term cold storage. Scientific Reports, 14, 30078. https://doi.org/10.1038/s41598-024-77871-z

Musio, B., Todisco, S., Antonicelli, M., Garino, C., Arlorio, M., Mastrorilli, P., Latronico, M., & Gallo, V. (2022). Non-Targeted NMR Method to Assess the Authenticity of Saffron and Trace the Agronomic Practices Applied for Its Production. Applied Sciences, 12(5), 2583. https://doi.org/10.3390/app12052583

Muthuramalingam, P., Krishnan, S. R., Pandian, S., Mareeswaran, N., Aruni, W., Pandian, S. K., & Ramesh, M. (2018). Global analysis of threonine metabolism genes unravel key players in rice to improve the abiotic stress tolerance. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-27703-8

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. https://bitcoin.org/bitcoin.pdf

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings of the twenty-first international conference on Machine learning, 78. https://doi.org/10.1145/1015330.1015435

Nyitrainé Sárdy, Á. D., Ladányi, M., Varga, Z., Szövényi, Á. P., & Matolcsi, R. (2022). The effect of grapevine variety and wine region on the primer parameters of wine based on 1H

NMR-spectroscopy and machine learning methods. Diversity, 14(2), 74. https://doi.org/10.3390/d14020074

Pais, I. P., Moreira, R., Semedo, J. N., Ramalho, J. C., Lidon, F. C., Coutinho, J., Maçãs, B., & Scotti-Campos, P. (2023). Wheat Crop under Waterlogging: Potential Soil and Plant Effects. In Plants (Vol. 12, Issue 1). MDPI. https://doi.org/10.3390/plants12010149

Pappalardo, L. (2022). Pomegranate fruit juice adulteration with apple juice: Detection by UV–visible spectroscopy combined with multivariate statistical analysis. Scientific Reports, 12, 5151. https://doi.org/10.1038/s41598-022-07979-7

Pascual García-Pérez, P. P., Becchi, L., Zhang, L., Rocchetti, G., & Lucini, L. (2024). Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity. Trends in Food Science & Technology, 147, 104481. https://doi.org/10.1016/j.tifs.2024.104481

Patelli, N., & Mandrioli, M. (2020).Blockchain technology and traceability in the agrifood industry. Journal of Food Science, 85(11), 3670–3678. https://doi.org/10.1111/1750-3841.15477

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. Computational Statistics and Data Analysis, 49(4), 974–997. https://doi.org/10.1016/j.csda.2004.06.015

Pérez-Toledo, M., González-Rodríguez, J., & Rodríguez-Méndez, M. L. (2022). Multivariate Chemometric Tools for Classification and Authentication of Extra Virgin Olive Oils Based on Volatile Fingerprinting. Foods, 11(3), 354. https://doi.org/10.3390/foods11030354

Picone, G. (2024). The 1H HR-NMR methods for the evaluation of the stability, quality, authenticity, and shelf life of foods. Encyclopedia, 4(4), 1617–1628. https://doi.org/10.3390/encyclopedia4040106

Pin, L., Sobolev, A. P., Testone, G., Scioli, G., Pinzari, F., Magnanimi, F., Colla, G., Cardarelli, M., & Giannino, D. (2025). Untargeted NMR Study of Metabolic Changes in Processing Tomato Treated with *Trichoderma atroviride* Under Open-Field Conditions and Exposed to Heatwave Temperatures. *Molecules*, *30*(1), 97. https://doi.org/10.3390/molecules30010097

Pollak, J., Mayonu, M., Jiang, L., & Wang, B. (2024). The development of machine learning approaches in two-dimensional NMR data interpretation for metabolomics applications. Analytical Biochemistry, 695, 115654. https://doi.org/10.1016/j.ab.2024.115654

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. In Computational and Structural Biotechnology Journal (Vol. 18, pp. 2818–2825). Elsevier B.V. https://doi.org/10.1016/j.csbj.2020.09.033

Powers, R. (2020). NMR Metabolomics and Statistical Modeling: Current Challenges and Future Directions. Metabolites, 10(8), 293. https://doi.org/10.3390/metabo10080293

Prechelt, L. (1998). Early stopping—but when? In Neural Networks: Tricks of the Trade (pp. 55–69). Springer. https://doi.org/10.1007/3-540-49430-8_3

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/widm.1301

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Rachoń, L., Cebulak, T., Krochmal-Marczak, B., Kapusta, I., Betlej, I., & Kiełtyka-Dadasiewicz, A. (2025). The Phenolic Acid Content in Wheat Depending on the Intensification of Cultivation Technology. Foods, 14(4). https://doi.org/10.3390/foods14040633

Radaelli, M., Scalabrin, E., Roman, M., Buffa, G., Griffante, I., & Capodaglio, G. (2024). Characterization of Ancient Cereals Cultivated by Intensive and Organic Procedures for Element Content. Molecules, 29(15). https://doi.org/10.3390/molecules29153645

Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo, M. Remedios Sillero-Denamiel, Variable selection for Naïve Bayes classification, Computers & Operations Research, Volume 135, 2021, 105456, ISSN 0305-0548, https://doi.org/10.1016/j.cor.2021.105456.

Ragone, R., Todisco, S., Triggiani, M., Pontrelli, S., Latronico, M., Mastrorilli, P., Intini, N., Ferroni, C., Musio, B., & Gallo, V. (2020). Development of a food class-discrimination system by non-targeted NMR analyses using different magnetic field strengths. Food Chemistry, 127339. https://doi.org/10.1016/j.foodchem.2020.127339

Ralli, E., & Spyros, A. (2023). A study of Greek Graviera cheese by NMR-based metabolomics. Molecules, 28(14), 5488. https://doi.org/10.3390/molecules28145488

Ramiro, J. L., Neo, A. G., Pérez-Palacios, T., Antequera, T., & Marcos, C. F. (2024). Machine learning-enabled fatty acid quantification and classification of pork from autochthonous breeds using low-field ^{1}H NMR spectroscopic data. Food Control, 166, 110753. https://doi.org/10.1016/j.foodcont.2024.110753

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. Analytica Chimica Acta, 885, 17–32. https://doi.org/10.1016/j.aca.2015.06.003

Rivera-Pérez, A., Romero-González, R., & Garrido Frenich, A. (2023). Untargeted ^{1}H NMR-based metabolomics and multi-technique data fusion: A promising combined approach for geographical and processing authentication of thyme by multivariate statistical analysis. Food Chemistry, 420, 136156. https://doi.org/10.1016/j.foodchem.2023.136156

Romano, G., Del Coco, L., Milano, F., Durante, M., Palombieri, S., Sestili, F., Visioni, A., Jilal, A., Fanizzi, F. P., & Laddomada, B. (2022). Phytochemical Profiling and Untargeted Metabolite Fingerprinting of the MEDWHEALTH Wheat, Barley and Lentil Wholemeal Flours. Foods, 11(24). https://doi.org/10.3390/foods11244070

Rose, T., Wilkinson, M., Lowe, C., Xu, J., Hughes, D., Hassall, K. L., Hassani-Pak, K., Amberkar, S., Noleto-Dias, C., Ward, J., & Heuer, S. (2022). Novel molecules and target genes for vegetative heat tolerance in wheat. Plant-Environment Interactions, 3(6), 264–289. https://doi.org/10.1002/pei3.10096

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. B. (2014). Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics, 10(3), 361–374. https://doi.org/10.1007/s11306-013-0598-6

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Santos, A. D. C., Fonseca, F. A., Lião, L. M., Alcantara, G. B., & Barison, A. (2015). High-resolution magic angle spinning nuclear magnetic resonance in foodstuff analysis. TrAC Trends in Analytical Chemistry, 73, 10–18. https://doi.org/10.1016/j.trac.2015.05.003

Savorani, F., Capozzi, F., Engelsen, S. B., Dell'Abate, M. T., & Sequi, P. (2009). Pomodoro di Pachino: An Authentication Study Using 1 H-NMR and Chemometrics – Protecting its P.G.I. European Certification (pp. 158–166). https://doi.org/10.1039/9781847559494-00158

Savorani, F., Tomasi, G., & Engelsen, S. B. (2009). 1H NMR-based chemometric analysis of tomato concentrates: A study of the effect of processing on the metabolite profile. Journal of Agricultural and Food Chemistry, 57(19), 8881–8889. https://doi.org/10.1021/jf901435u

Schievano, E., Peggion, E., & Mammi, S. (2010). ^{1}H NMR spectra of chloroform extracts of honey for chemometric determination of its botanical origin. Journal of Agricultural and Food Chemistry, 58(1), 57–65. https://doi.org/10.1021/jf902973r

Selamat, J., Rozani, N. A. A., & Murugesu, S. (2021). Application of the metabolomics approach in food authentication. Molecules, 26(24), 7565. https://doi.org/10.3390/molecules26247565

Shamanin, V. P., Tekin-Cakmak, Z. H., Gordeeva, E. I., Karasu, S., Pototskaya, I., Chursin, A. S., Pozherukova, V. E., Ozulku, G., Morgounov, A. I., Sagdic, O., & Koksel, H. (2022). Antioxidant Capacity and Profiles of Phenolic Acids in Various Genotypes of Purple Wheat. Foods, 11(16). https://doi.org/10.3390/foods11162515

Shewry, P. R., Corol, D. I., Jones, H. D., Beale, M. H., & Ward, J. L. (2017). Defining genetic and chemical diversity in wheat grain by 1H-NMR spectroscopy of polar metabolites. Molecular Nutrition and Food Research, 61(7). https://doi.org/10.1002/mnfr.201600807

Sobolev, A. P., Ingallina, C., Spano, M., Di Matteo, G., & Mannina, L. (2022). NMR-Based Approaches in the Study of Foods. Molecules, 27(22), 7906. https://doi.org/10.3390/molecules27227906

Sokal, Robert & Rohlf, F.. (2013). Biometry : the principles and practice of statistics in biological research / Robert R. Sokal and F. James Rohlf.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Song, E. H., Kim, H. J., Jeong, J., Chung, H. J., Kim, H. Y., Bang, E., & Hong, Y. S. (2016). A 1H HR-MAS NMR-based metabolomic study for metabolic characterization of rice grain from various Oryza sativa L. cultivars. Journal of Agricultural and Food Chemistry, 64(15), 3009–3016. https://doi.org/10.1021/acs.jafc.5b05667

Spanic, V., Duvnjak, J., Hefer, D., & D'Auria, J. C. (2025). Changes in Metabolites Produced in Wheat Plants Against Water-Deficit Stress. Plants, 14(1). https://doi.org/10.3390/plants14010010

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. In Source: Journal of the American Statistical Association (Vol. 69, Issue 347).

Stoyanova, R., & Brown, T. R. (2001). NMR spectral quantitation by principal component analysis. NMR in Biomedicine, 14(4), 271–277. https://doi.org/10.1002/nbm.700

Sundekilde, U. K., Poulsen, N. A., Larsen, L. B., & Bertram, H. C. (2013). Nuclear magnetic resonance metabonomics reveals strong association between milk metabolites and somatic cell count in bovine milk. Journal of Dairy Science, 96(1), 290–299. https://doi.org/10.3168/jds.2012-5819

Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics, 8(1), 3–16. https://doi.org/10.1007/s11306-011-0330-3

Tian, F. (2016). An agri-food supply chain traceability system for China based on RFID & blockchain technology. 13th International Conference on Service Systems and Service Management (ICSSSM), 1–6. https://doi.org/10.1109/ICSSSM.2016.7538424

Tian, F. (2017). A supply chain traceability system for food safety based on HACCP, blockchain & Internet of things. 2017 International Conference on Service Systems and Service Management, 1–6. https://doi.org/10.1109/ICSSSM.2017.7996119

Tomassi, E., Arouna, N., Brasca, M., Silvetti, T., de Pascale, S., Troise, A. D., Scaloni, A., & Pucci, L. (2025). Fermentation of Whole-Wheat Using Different Combinations of Lactic Acid Bacteria and Yeast: Impact on In Vitro and Ex Vivo Antioxidant Activity. Foods, 14(3). https://doi.org/10.3390/foods14030421

Tripoli, M., & Schmidhuber, J. (2018). Emerging opportunities for the application of blockchain in the agri-food industry. FAO and ITC. http://www.fao.org/3/CA2906EN/ca2906en.pdf

Truzzi, E., Marchetti, L., Benvenuti, S., Ferroni, A., Rossi, M.C., & Bertelli, D. (2021). Novel strategy for the recognition of adulterant vegetable oils in essential oils commonly used in food industries by applying ^13C NMR spectroscopy. Journal of Agricultural and Food Chemistry, 69(29), 8276–8286. https://doi.org/10.1021/acs.jafc.1c02279

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics: A Journal of the Chemometrics Society, 16(3), 119–128. https://doi.org/10.1002/cem.695

Tsermoula, P., Kristensen, N. B., Mobaraki, N., Engelsen, S. B., & Khakimov, B. (2024). Efficient Quantification of Milk Metabolites from 1H NMR Spectra Using the Signature Mapping (SigMa) Approach: Chemical Shift Library Development for Cows' Milk and Colostrum. Analytical Chemistry, 96(5), 1861–1871. https://doi.org/10.1021/acs.analchem.3c03449

Utpott, M., Rodrigues, E., Rios, A. de O., Mercali, G. D., & Flôres, S. H. (2022). Metabolomics: An analytical technique for food processing evaluation. Food Chemistry, 366, 130685. https://doi.org/10.1016/j.foodchem.2021.130685

Valentini, M., Ritota, M., Cafiero, C., Cozzolino, S., Leita, L., & Sequi, P. (2011). The HRMAS–NMR tool in foodstuff characterisation. Magnetic Resonance in Chemistry, 49(S1), S121–S125. https://doi.org/10.1002/mrc.2826

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC Genomics, 7(1), 142. https://doi.org/10.1186/1471-2164-7-142

Varavallo, G., Caragnano, G., Bertone, F., Vernetti-Prot, L., & Terzo, O. (2022). Traceability Platform Based on Green Blockchain: An Application Case Study in Dairy Supply Chain. Sustainability, 14(6), 3321. https://doi.org/10.3390/su14063321

Vigli, G., Philippidis, A., Spyros, A., & Dais, P. (2003). Classification of edible oils by employing 31P and 1H NMR spectroscopy in combination with multivariate statistical analysis. A comparative study. Journal of Agricultural and Food Chemistry, 51(19), 5715–5722. https://doi.org/10.1021/jf034225+

Vu, T., Siemek, P., Bhinderwala, F., Xu, Y., & Powers, R. (2019). Evaluation of Multivariate Classification Models for Analyzing NMR Metabolomics Data. Journal of Proteome Research, 18(9), 3282–3294. https://doi.org/10.1021/acs.jproteome.9b00227

Vujović, Ž. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M., & van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. Metabolomics, 4, 81–89. https://doi.org/10.1007/s11306-007-0099-6

Wishart, D. S. (2008). Metabolomics: Applications to food science and nutrition research. Trends in Food Science & Technology, 19(9), 482–493. https://doi.org/10.1016/j.tifs.2008.03.003

Wishart, D. S., Cheng, L. L., Copié, V., Edison, A. S., Eghbalnia, H. R., Hoch, J. C., Gouveia, G. J., Pathmasiri, W., Powers, R., Schock, T. B., Sumner, L. W., & Uchimiya, M. (2022). NMR and metabolomics—A roadmap for the future. Metabolites, 12(8), 678. https://doi.org/10.3390/metabo12080678

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. Current Metabolomics, 1(1), 92–107. https://doi.org/10.2174/2213235X11301010092

Wu W, Zhang L, Zheng X, Huang Q, Farag MA, Zhu R, Zhao C. Emerging applications of metabolomics in food science and future trends. Food Chem X. 2022 Nov 9;16:100500. doi: 10.1016/j.fochx.2022.100500. PMID: 36519105; PMCID: PMC9743159

Yakar, Y., & Karadağ, K. (2022). Identifying olive oil fraud and adulteration using machine learning algorithms. Química Nova, 45(10), 1245–1250. https://doi.org/10.21577/0100-4042.20170948

Yan, F., Ma, J., Peng, M., Xi, C., Chang, S., Yang, Y., Tian, S., Zhou, B., & Liu, T. (2024). Lactic acid induced defense responses in tobacco against Phytophthora nicotianae. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-60037-2

Yata, K., & Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. Journal of Multivariate Analysis, 105(1), 193–215. https://doi.org/10.1016/j.jmva.2011.09.002

Zhang, Y., et al. (2021). PLS-DA score plots derived from the 1H-NMR spectra of soybean samples. Recuperato da https://www.researchgate.net/figure/PLS-DA-score-plots-A-derived-from-the-1-H-NMR-spectra-of-soybean-samples-for_fig1_349378284

Zhao, P., Gu, S., Han, C., Lu, Y., Ma, C., Tian, J., Bi, J., Deng, Z., Wang, Q., & Xu, Q. (2021). Targeted and Untargeted Metabolomics Profiling of Wheat Reveals Amino Acids Increase Resistance to Fusarium Head Blight. Frontiers in Plant Science, 12. https://doi.org/10.3389/fpls.2021.762605

Zheng, S., Hao, Y., Fan, S., Cai, J., Chen, W., Li, X., & Zhu, X. (2021). Metabolomic and Transcriptomic Profiling Provide Novel Insights into Fruit Ripening and Ripening Disorder Caused by 1-MCP Treatments in Papaya. International Journal of Molecular Sciences, 22(2), 916. https://doi.org/10.3390/ijms22020916

Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2018). An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. In 2017 IEEE International Congress on Big Data (pp. 557–564). https://doi.org/10.1109/BigDataCongress.2017.85

Zhou, Z.-H. (2012). Ensemble Methods: Foundations and Algorithms. CRC Press.

Capitolo 7 Appendice

Nella figura seguente si mostra un estratto dello script utilizzato per la lettura dei dati contenuti nella matrice dei buckets spettrali e la successiva standardizzazione.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
import joblib
import os

# Caricamento del dataset
df = pd.read_csv("/content/MATRICE.csv")

# Controllo valori mancanti
missing_values = df.isnull().sum().sum()
if missing_values > 0:
    print(f" Attenzione: ci sono {missing_values} valori mancanti nel
dataset!")
else:
    print(" Nessun valore mancante nel dataset.")

# Separazione features (X) e target (y)
X = df.iloc[:, 1:-1]  # esclude prima colonna (Replica) e ultima
(Class)
y = df['Class']

# Standardizzazione dei dati
scaler = StandardScaler()
X_scaled = pd.DataFrame(scaler.fit_transform(X), columns=X.columns)

#  Mostra tabella descrittiva dei dati standardizzati
print("\n Statistiche dei Dati Standardizzati:")
display(X_scaled.describe().T)

#  Salvataggio statistiche descrittive in CSV
X_scaled.describe().T.to_csv("/content/Statistiche_Dati_Standardizzati.
csv")
print(" Statistiche dati standardizzati salvate in:
Statistiche_Dati_Standardizzati.csv")
```

Figura 7.1 – Script per la lettura dei dati della matrice dei buckets e la standardizzazione Z-score

Nella figura seguente si mostra un estratto dello script dedicato all'esecuzione della PCA.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import itertools
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from mpl_toolkits.mplot3d import Axes3D
import pickle
import os

# 1. Caricamento e Preprocessamento dei Dati
def load_and_preprocess_data(file_path):
    data = pd.read_csv(file_path)
    X = data.iloc[:, 1:-1]
    y = data["Class"]
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    os.makedirs("/content/saved_models", exist_ok=True)
    with open("/content/saved_models/scaler.pkl", "wb") as f:
        pickle.dump(scaler, f)
    print("Scaler salvato in: /content/saved_models/scaler.pkl")

    return X, y, X_scaled8

# 2. Scelta Automatica del Numero di PC
def select_optimal_pcs(X_scaled):
    n_components = min(X_scaled.shape[0], X_scaled.shape[1])
    pca = PCA(n_components=n_components)
    pca.fit(X_scaled)

    cumulative_variance = np.cumsum(pca.explained_variance_ratio_)
    num_pc_95 = sum(cumulative_variance <= 0.95) + 1
    num_pc_kaiser = sum(pca.explained_variance_ > 1)

    # Elbow Plot
    plt.figure(figsize=(8, 5))
    plt.plot(range(1, len(cumulative_variance) + 1),
cumulative_variance, marker='o')
    plt.axhline(y=0.95, color='red', linestyle='--', label='95%
varianza spiegata')
    plt.title('Elbow Plot - Varianza Cumulativa')
    plt.xlabel('Numero di Componenti Principali')
    plt.ylabel('Varianza Cumulativa')
    plt.grid(True)
    plt.legend()
    plt.tight_layout()
    plt.savefig("/content/Elbow_Plot_PCA.png")
    plt.show()
    print("Elbow plot salvato in: /content/Elbow_Plot_PCA.png")

    # Broken Stick Criterion
    broken_stick = np.array([sum(1 / np.arange(i, n_components + 1))
for i in range(1, n_components + 1)])
```

```
    broken_stick = broken_stick / sum(broken_stick)
    num_pc_broken_stick = sum(pca.explained_variance_ratio_ >
broken_stick[:len(pca.explained_variance_ratio_)])

    print("\n Risultati della selezione delle componenti principali
(PC):")
    print(f"  PC selezionati con varianza ≥ 95%: {num_pc_95}")
    print(f"  PC selezionati con criterio di Kaiser (λ > 1):
{num_pc_kaiser}")
    print(f"  PC selezionati con Broken Stick: {num_pc_broken_stick}")
    if num_pc_95 <= 10:
        selected_pc = num_pc_95
    elif num_pc_kaiser <= 10:
        selected_pc = min(num_pc_95, num_pc_kaiser)
    elif abs(num_pc_broken_stick - num_pc_95) <= 1:
        selected_pc = num_pc_broken_stick
    else:
        selected_pc = num_pc_95

    print(f"\nNumero di componenti scelto automaticamente:
{selected_pc}")
    return selected_pc

# 3. Esecuzione della PCA
def perform_pca(X_scaled, selected_pcs, y):
    pca = PCA(n_components=selected_pcs)
    X_pca = pca.fit_transform(X_scaled)

    autovalori = pca.explained_variance_
    autovettori = pca.components_

    df_loadings = pd.DataFrame(
        pca.components_,
        columns=[f"Var{i+1}" for i in range(X_scaled.shape[1])],
        index=[f"PC{i+1}" for i in range(selected_pcs)]
    )

    df_scores = pd.DataFrame(X_pca, columns=[f"PC{i+1}" for i in
range(selected_pcs)])
    df_scores["Class"] = y.values

    df_contribution_matrix = (df_loadings ** 2)
    df_contribution_matrix =
df_contribution_matrix.div(df_contribution_matrix.sum(axis=1), axis=0)
* 100
    df_contribution_matrix = df_contribution_matrix.transpose()
    df_contribution_matrix.to_csv("/content/PCA_Contributi_Variabili_St
ile_XLSTAT.csv")
    df_autovalori = pd.DataFrame({"PC": [f"PC{i+1}" for i in
range(selected_pcs)], "Autovalori": autovalori})
    df_autovettori = pd.DataFrame(autovettori, columns=[f"Var{i+1}" for
i in range(df_loadings.shape[1])], index=[f"PC{i+1}" for i in
range(selected_pcs)])

    df_autovalori.to_csv(f"PCA_Autovalori_{selected_pcs}PC.csv",
index=False)
    df_autovettori.to_csv(f"PCA_Autovettori_{selected_pcs}PC.csv")
    df_loadings.to_csv(f"PCA_Loadings_{selected_pcs}PC.csv")
```

```
    df_scores.to_csv(f"PCA_Scores_{selected_pcs}PC.csv", index=False)

    with open("/content/saved_models/pca_model.pkl", "wb") as f:
        pickle.dump(pca, f)

    return autovalori, autovettori, df_loadings, df_scores,
df_contribution_matrix
```

Figura 7.2 - Script per l'esecuzione della PCA

Nella figura seguente è riportato un estratto dello script utilizzato per l'addestramento, l'ottimizzazione degli iperparametri e il salvataggio dei modelli di classificazione in presenza di dati affetti da rumore gaussiano. Lo script comprende la valutazione delle performance predittive mediante cross-validazione stratificata a n fold.

```
import pandas as pd
import numpy as np
import pickle
import glob
import os
import json
from sklearn.model_selection import StratifiedKFold, GridSearchCV,
cross_val_score
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.inspection import permutation_importance
from sklearn.metrics import classification_report, confusion_matrix
import warnings
warnings.filterwarnings("ignore")

# Impostazioni iniziali
use_noise = True  #
noise_std = 0.1
k = 5  # numero di fold

# Step 1: Imposta seed
np.random.seed(42)
os.environ['PYTHONHASHSEED'] = str(42)

# Step 2: Recupera numero di PC
def get_selected_pcs():
    file = glob.glob("/content/PCA_Loadings_*PC.csv")[0]
    return int(file.split("_")[-1].replace("PC.csv", ""))

# Step 3: Carica dati PCA e classi
def load_data(selected_pcs):
    df = pd.read_csv(f"/content/PCA_Scores_{selected_pcs}PC.csv")
```

```
    y = df['Class']
    X = df.drop(columns=['Class'])
    le = LabelEncoder()
    y_enc = le.fit_transform(y)
    os.makedirs("/content/saved_models", exist_ok=True)
    with open("/content/saved_models/label_encoder.pkl", "wb") as f:
        pickle.dump(le, f)
    with open("/content/saved_models/class_mapping.json", "w") as f:
        json.dump(dict(enumerate(le.classes_)), f, indent=4)
    return X, y_enc

# Step 4: Aggiunta rumore
def add_noise(X):
    noise = np.random.normal(0, noise_std, X.shape)
    return X + noise

# Step 5: Definizione modelli e griglie
models = {
    "Random Forest": (RandomForestClassifier(random_state=42), {
        "n_estimators": [100, 200],
        "max_depth": [None, 10, 20]
    }),
    "SVM": (SVC(probability=True, random_state=42), {
        "C": [0.1, 1, 10],
        "kernel": ["linear", "rbf"]
    }),
    "XGBoost": (XGBClassifier(eval_metric='mlogloss', random_state=42),
{
        "n_estimators": [100, 200],
        "max_depth": [3, 6],
        "learning_rate": [0.01, 0.1]
    }),
    "k-NN": (KNeighborsClassifier(), {
        "n_neighbors": [3, 5, 7],
        "metric": ["euclidean", "manhattan"]
    }),
    "Naive Bayes": (GaussianNB(), {}),
    "Decision Tree": (DecisionTreeClassifier(random_state=42), {
        "max_depth": [None, 10, 20]
    }),
    "Logistic Regression": (LogisticRegression(max_iter=1000,
random_state=42), {
        "C": [0.1, 1, 10]
    })
}

# Step 6: Training + salvataggi
cv = StratifiedKFold(n_splits=k, shuffle=True, random_state=42)
selected_pcs = get_selected_pcs()
X, y = load_data(selected_pcs)
if use_noise:
    X = add_noise(X)

results = []
best_models = {}

for name, (model, params) in models.items():
    print(f"\n GridSearchCV - {name}")
```

```
    if params:
        gs = GridSearchCV(model, params, cv=cv, scoring='accuracy',
n_jobs=-1)
        gs.fit(X, y)
        best = gs.best_estimator_
        best_params = gs.best_params_
        best_score = gs.best_score_
    else:
        best = model.fit(X, y)
        best_score = cross_val_score(best, X, y, cv=cv,
scoring='accuracy').mean()
        best_params = "default"
    acc_std = cross_val_score(best, X, y, cv=cv).std()

    results.append({
        "Modello": name,
        "Accuracy (mean)": round(best_score, 4),
        "Accuracy (std)": round(acc_std, 4),
        "Nota": nota,
        "Parametri ottimali": best_params
    })

    best_models[name] = best

    # Salva modello
    with open(f"/content/saved_models/GS_FINAL_{name.replace(' ',
'_')}.pkl", "wb") as f:
        pickle.dump(best, f)

    # Salva MDA
    perm = permutation_importance(best, X, y, n_repeats=10,
random_state=42)
    df_imp = pd.DataFrame({
        "Feature": [f"PC{i+1}" for i in range(X.shape[1])],
        "Mean Decrease Accuracy": perm.importances_mean
    }).sort_values(by="Mean Decrease Accuracy", ascending=False)
    df_imp.to_csv(f"/content/MDA_GS_FINAL_{name.replace(' ',
'_')}.csv", index=False)

    # Classification Report
    y_pred = best.predict(X)
    with open(f"/content/Report_{name.replace(' ', '_')}.txt", "w") as
f:
        f.write(classification_report(y, y_pred))
        f.write("\nConfusion Matrix:\n")
        f.write(str(confusion_matrix(y, y_pred)))
…
pd.DataFrame(results).to_csv("/content/Results_GS_FINAL.csv",
index=False)
```

Figura 7.3 - Script per l'ottimizzazione degli iperparametri, l'addestramento dei modelli e la selezione delle componenti discriminanti

Nella figura seguente è presentato un estratto dello script impiegato per la validazione dei modelli predittivi su dati non precedentemente utilizzati in fase di addestramento.

```
import os
import pickle
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import cross_val_predict, StratifiedKFold
from sklearn.metrics import (accuracy_score, precision_score,
recall_score,
                             f1_score, confusion_matrix,
ConfusionMatrixDisplay)
from sklearn.preprocessing import LabelEncoder
import glob
import warnings
warnings.filterwarnings("ignore")

# Impostazioni
SEED = 42
CV = StratifiedKFold(n_splits=5, shuffle=True, random_state=SEED)
MODEL_DIR = "/content/saved_models"

# Lista dei modelli da caricare
model_names = [
    "Random_Forest",
    "SVM",
    "XGBoost",
    "k-NN",
    "Naive_Bayes",
    "Decision_Tree",
    "Logistic_Regression"
]

# Caricamento dati PCA (con classi)
def load_pca_data():
    file = glob.glob("/content/PCA_Scores_*PC.csv")[0]
    df = pd.read_csv(file)
    y = df['Class']
    X = df.drop(columns=['Class'])
    le = LabelEncoder()
    y_enc = le.fit_transform(y)
    return X, y_enc, le

# Funzione per calcolo metriche

def calculate_metrics(y_true, y_pred):
    return {
        "Accuracy": accuracy_score(y_true, y_pred),
        "F1_macro": f1_score(y_true, y_pred, average='macro'),
        "F1_weighted": f1_score(y_true, y_pred, average='weighted'),
        "Precision_macro": precision_score(y_true, y_pred,
average='macro'),
        "Recall_macro": recall_score(y_true, y_pred, average='macro')
    }

# MAIN
X, y, le = load_pca_data()
results = []
```

```
for name in model_names:
    path = os.path.join(MODEL_DIR, f"GS_FINAL_{name}.pkl")
    print(f"\n Validazione - {name.replace('_', ' ')}")

    with open(path, 'rb') as f:
        model = pickle.load(f)

    y_pred = cross_val_predict(model, X, y, cv=CV)
    metrics = calculate_metrics(y, y_pred)
    metrics["Modello"] = name.replace("_", " ")
    results.append(metrics)


# Salvataggio risultati
results_df = pd.DataFrame(results)
results_df = results_df[["Modello", "Accuracy", "F1_macro",
"F1_weighted", "Precision_macro", "Recall_macro"]]
results_df.to_csv("/content/Metriche_Modelli_ConRumore.csv",
index=False)
print("\n Risultati salvati in: Metriche_Modelli_ConRumore.csv")
print(results_df)
```

Figura 7.4 - Script per la valutazione comparativa delle prestazioni predittive su dati non precedentemente utilizzati in fase di addestramento

Nella figura seguente è riportato un estratto dello script sviluppato per combinare i contributi percentuali delle variabili (derivanti dall'analisi PCA) con l'importanza delle componenti principali, calcolata tramite *permutation importance* a partire dai modelli predittivi. Il risultato di questa combinazione è un file finale che riporta l'importanza pesata di ciascuna variabile, distinta per ciascun modello considerato.

```
import pandas as pd
import numpy as np
import glob
import os

# 1. Rileva automaticamente il file dei contributi PCA in stile XLSTAT
def get_selected_pcs():
    contrib_file =
glob.glob("/content/PCA_Contributi_Variabili_Stile_XLSTAT.csv")
    if not contrib_file:
        raise FileNotFoundError("Nessun file
PCA_Contributi_Variabili_Stile_XLSTAT.csv trovato.")
    print(f"File contributi (XLSTAT) trovato: {contrib_file[0]}")
    return contrib_file[0]

# 2. Carica i contributi delle variabili (righe = variabili, colonne =
PC)
def load_variable_contributions(contrib_file):
    df = pd.read_csv(contrib_file, header=0)
    df = df.set_index(df.columns[0])
    df = df.apply(pd.to_numeric, errors='coerce').fillna(0)
    return df
```

```
# 3. Carica tutte le MDA dei modelli
def load_model_mda(pcs_da_usare):
    mda_dict = {}
    files = glob.glob("/content/MDA_GS_FINAL_*.csv")
    for path in files:
        model_name = path.split("MDA_NOISE_CV_")[-1].replace(".csv",
"").replace("_", " ")
        df = pd.read_csv(path)
        df.columns = ['PC', 'MDA']
        df['PC'] = df['PC'].str.strip()
        df = df[df['PC'].isin(pcs_da_usare)]
        if df.empty:
            print(f"{model_name}: nessuna PC corrispondente. Saltato.")
            continue
        df['MDA'] = df['MDA'] / df['MDA'].sum()
        mda_dict[model_name] = df.set_index('PC')['MDA']
        print(f"MDA caricata per: {model_name}")
    return mda_dict

# 4. Calcola l'importanza finale delle variabili usando MDA
def compute_variable_importance(contrib_df, model_mda_dict):
    risultati = []
    for modello, mda_series in model_mda_dict.items():
        imp_tot = pd.Series(0, index=contrib_df.index, dtype=float)
        for pc, weight in mda_series.items():
            if pc in contrib_df.columns:
                imp_tot += contrib_df[pc] * weight
        df_model = pd.DataFrame({
            'Variabile': imp_tot.index,
            'Importanza_MDA': imp_tot.values,
            'Modello': modello
        }).sort_values(by='Importanza_MDA',
ascending=False).reset_index(drop=True)
        risultati.append(df_model)
    return pd.concat(risultati, ignore_index=True)

# 5. Main
def main():
    contrib_file = get_selected_pcs()
    contrib_df = load_variable_contributions(contrib_file)
    pcs_da_usare = contrib_df.columns.tolist()
    model_mda = load_model_mda(pcs_da_usare)
    df_finale = compute_variable_importance(contrib_df, model_mda)

    if not df_finale.empty:
        output_path = "/content/Importanza_Variabili_da_MDA.csv"
        df_finale.to_csv(output_path, index=False)
        print(f"\n File finale salvato in: {output_path}")
        print(df_finale.groupby("Modello").head(5))
    else:
        print("Nessun risultato da salvare.")

main()
```

Figura 7.5 - Script per il calcolo dell'importanza finale delle variabili combinando i contributi percentuali della PCA con i valori di permutaion importance ottenuti dai modelli predittivi

Nella figura seguente viene riportato un estratto dello script sviluppato per il test di predizione dei modelli applicato a dati del set esterno.

```
import pandas as pd
import numpy as np
import joblib
import pickle
from sklearn.metrics import accuracy_score, f1_score,
classification_report, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
import os

# === 1. Impostazioni === #
data_path = "/content/FarzatiFrumenti_tot.csv"
scaler_path = "/content/saved_models/StandardScaler_model.pkl"
pca_path = "/content/saved_models/pca_model.pkl"
label_encoder_path = "/content/saved_models/label_encoder.pkl"
output_dir = "/content/predictions_results"
os.makedirs(output_dir, exist_ok=True)

# === 2. Caricamento dati === #
df = pd.read_csv(data_path)
repliche = df["Replica"] if "Replica" in df.columns else df["replica"]
X = df.iloc[:, 1:-1]
y_true = df["Class"]

# === 3. Caricamento scaler, PCA e LabelEncoder === #
scaler = joblib.load(scaler_path)
X_scaled = scaler.transform(X)

with open(pca_path, "rb") as f:
    pca = pickle.load(f)
X_pca = pca.transform(X_scaled)

with open(label_encoder_path, "rb") as f:
    label_encoder = pickle.load(f)

y_true_encoded = label_encoder.transform(y_true)

# === 4. Loop su tutti i modelli === #
model_names = ["Random_Forest", "SVM", "XGBoost", "k-NN",
"Naive_Bayes", "Decision_Tree", "Logistic_Regression"]

for name in model_names:
    print(f"\n📌 Valutazione modello: {name}")
    model_path = f"/content/saved_models/GS_FINAL_{name}.pkl"
    with open(model_path, "rb") as f:
        model = pickle.load(f)

    y_pred = model.predict(X_pca)
    y_pred_decoded = label_encoder.inverse_transform(y_pred)

    accuracy = accuracy_score(y_true_encoded, y_pred)
    f1_macro = f1_score(y_true_encoded, y_pred, average="macro")
    f1_weighted = f1_score(y_true_encoded, y_pred, average="weighted")
```

```
    print(classification_report(y_true, y_pred_decoded))
    print(f"Accuracy: {accuracy:.4f} | F1 Macro: {f1_macro:.4f} | F1
Weighted: {f1_weighted:.4f}")

    # Salvataggio predizioni
    df_result = pd.DataFrame({
        "Replica": repliche,
        "Classe Reale": y_true,
        "Classe Predetta": y_pred_decoded
    })
    df_result.to_csv(f"{output_dir}/Predizioni_{name}.csv",
index=False)

    # Confusion Matrix
    cm = confusion_matrix(y_true, y_pred_decoded,
labels=label_encoder.classes_)
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",
                xticklabels=label_encoder.classes_,
                yticklabels=label_encoder.classes_)
    plt.xlabel("Classe Predetta")
    plt.ylabel("Classe Reale")
    plt.title(f"Matrice di Confusione - {name}")
    plt.tight_layout()
    plt.savefig(f"{output_dir}/ConfusionMatrix_{name}.png")
    plt.close()
```

Figura 7.6 - Script per la valutazione delle performance predittive con set di dati esterno

```
# 1. Installazione libreria
!pip install nmrglue

# 2. Importazione i moduli
import nmrglue as ng
import numpy as np
import matplotlib.pyplot as plt
import json, hashlib, os
from datetime import datetime
from google.colab import files
import zipfile

# 3. Caricamento file ZIP Bruker
uploaded = files.upload()
for filename in uploaded.keys():
    zip_path = filename
    extract_dir = filename.replace(".zip", "")
    with zipfile.ZipFile(zip_path, 'r') as zip_ref:
        zip_ref.extractall(extract_dir)

# 4. Lettura del FID Bruker (dati grezzi)
dic, fid = ng.bruker.read(extract_dir)
```

```
# 5. Estrazione metadati da acqus
acqus = dic.get('acqus', {})

def get_param(key, default="unknown"):
    return acqus.get(key) or acqus.get(f"${key}") or default

pulse_program = get_param("PULPROG").replace(";", "").strip()
raw_date = get_param("DATE", "unknown")
try:
    raw_date = float(raw_date)
    date = datetime.utcfromtimestamp(raw_date).isoformat()
except:
    date = str(raw_date)

try:
    temperature = float(get_param("TE", 298.0))
except:
    temperature = "unknown"

# 6. Costruizione JSON con dati FID grezzi
spettro_json = {
    "sample_id": "id campione",
    "matrix": "matrice",
    "instrument": "av400",
    "operator": "operatore",
    "date": date,
    "temperature": temperature,
    "pulse_program": pulse_program,
    "fid_real": np.real(fid).tolist(),
    "fid_imag": np.imag(fid).tolist(),
    "is_domain": "time",
    "timestamp": datetime.utcnow().isoformat()
}

# 7. Salvataggio JSON
json_filename = "raw_fid.json"
with open(json_filename, "w") as f:
    json.dump(spettro_json, f, indent=4)

# 8. Calcolo hash SHA256
with open(json_filename, "rb") as f:
    content = f.read()
    hash_digest = hashlib.sha256(content).hexdigest()

print("Hash SHA256 generato per la blockchain:")
print(hash_digest)
```

Figura 7.7 - Script per serializzazione e generazione hash da FID Bruker