UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

SPETTROSCOPIA NEL VICINO INFRAROSSO PER LO SMART AGRIFOOD: REVISIONE CRITICA DEI METODI DI ANALISI E PROSPETTIVE FUTURE

Relatrice: Giulia Cisotto

Correlatore: Leonardo Badia

Laureando: Matteo Bergamin

ANNO ACCADEMICO 2021/2022

Data di laurea 16/11/2022

Abstract

Il crescente fabbisogno mondiale di cibo e la necessità di controlli serrati per contrastare le frodi alimentari hanno determinato l'esigenza di sviluppare nuove tecniche per l'analisi di prodotti agroalimentari. Le tecniche tradizionali che utilizzano la cromatografia liquida, la spettrometria di massa o tecniche basate sullo studio del DNA sono costose, richiedono reagenti chimici e molto tempo per l'analisi. Numerosi studi recenti applicano la spettrometria nel vicino infrarosso come metodo alternativo alle tecniche di analisi tradizionali. La spettrometria NIR è una tecnica di analisi che studia la reazione di alcune molecole all'esposizione di onde elettromagnetiche nella regione dell'infrarosso. Grazie ad essa è possibile estrarre informazioni dai prodotti in maniera rapida, non invasiva e a basso costo. Questa tesi si inserisce nell'ambito di una collaborazione tra l'Università degli Studi di Padova, il Dipartimento di Informatica dell'Università degli Studi di Milano-Bicocca e l'azienda Seletech Engineering Srl di Brughiero (Milano). In questa tesi attraverso una revisione critica della letteratura si studia l'evoluzione della tecnica con una particolare attenzione alle caratteristiche estratte dallo spettro elettromagnetico e alle tecniche di pre-elaborazione dei dati e di analisi multivariata.

Indice

Elenco degli acronimi

Acd	*Acidità*
Amd	*Amido*
BrimA	*Indice di misurazione tra dolcezza e acidità*
Cal	*Calcare*
CC	*Corretta classificazione*
Co	*Carbonio organic*
Cp	*Proteina cruda*
Crb	*Carbonile*
C/N	*Rapporto carbonio azoto*
CNN	*Convolutional Neural Network*
ELM	*Extreme Learning Machine*
Fru	*Fruttosio*
FCI	*Indice del colore del frutto*
FW	*Peso del frutto o campione*
Glu	*Glucosio*
Gr	*Grasso*
HSI	*Hyperspectral imaging*
Mel	*Melanina*
MLR	*Multiple Linear Regression*
MI	*Indice di maturità*
MSC	*Multiplicative Scatter Correction*
NIRS	*Near Infrared Spectroscopy*
Nt	*Azoto totale*
PLS	*Partial Least Square*
PLS-DA	*Partial Least Square for discriminant analysis*
Pr	*Proteine*
R^2	*Coefficiente di determinazione*
R_p^2	*Coefficiente di determinazione dei valori predetti*
R_c^2	*Coefficiente di determinazione di calibrazione*
R_{cv}^2	*Coefficiente di determinazione di cross-validazione*
RMSE	*Radice dell'errore quadratico medio*
RMSEP	*Radice dell'errore quadratico medio dei valori predetti*
RMSECV	*Radice dell'errore quadratico medio di cross-validazione*
RPD	*Deviazione dei residui predetti*
Su	*Zucchero*
SNV	*Standard Normal Variate*
SVM	*Support Vector Machine*
S-ELM	*Subagging ELM*
SSAE-AT-ELM	*Stacked Sparse Autoencoder affine transformation with Extreme Learning Machine*
Sn	*Sensibilità*
Sp	*Specificità*
SAD	*Software di acquisizione dati*
SAS	*Software per l'analisi dei dati*
S-G	*Savitzky-Golay*
SICMA	*Soft Independent Modeling of Class Analogies*
SVD	*Singular Value Decomposition*
TSS	*Totale di solidi solubili*
TA	*Acido citrico*
TSS: TA	*Rapporto TSS: TA*
Um	*Umidità*
VWCNN	*Variable Weight Convolutional Neural Network*
V-C	*Vitamina C*

Capitolo 1

Introduzione

Negli ultimi anni il settore agroalimentare è in continua espansione ma deve far fronte a numerose problematiche come la sovrappopolazione mondiale e la necessità di controlli serrati per evitare le frodi alimentari. L'organizzazione delle Nazioni Unite per l'alimentazione e l'agricoltura (FAO) afferma che la produzione di cibo mondiale deve aumentare fino al 70% entro il 2050 per venire in contro al fabbisogno mondiale [41]. Sempre di più sono anche i casi di frodi alimentari che si sono verificati negli ultimi anni. Nel 2008, diverse aziende cinesi furono coinvolte in uno scandalo riguardante l'adulterazione del latte con la melanina che causarono un'insufficienza renale ed infezioni del tratto urinario a circa 300.000 bambini [25].

I metodi di analisi tradizionali utilizzano la cromatografia liquida, la spettrometria di massa o tecniche basate sullo studio del DNA. L'applicazione di questi metodi richiede molto tempo, sono a bassa efficienza, hanno un alto costo per misurazione, richiedono personale qualificato e comportano la distruzione del campione [31].

L'ultimo decennio ha visto uno sviluppo senza precedenti della tecnologia, specialmente per quanto riguarda *l'Intelligenza Artificiale* [43], la *sensoristica intelligente* [44] e l'*Internet of Things* (IoT) [42]. La spettrometria nel vicino infrarosso applicata al settore agroalimentare si presenta come una tecnica alternativa per l'analisi quantitativa e qualitativa rispetto alle tecniche tradizionali da laboratorio. La spettrometria NIR a discapito delle tecniche di analisi tradizionali offre il vantaggio di poter effettuare misurazioni a basso costo in modo rapido e non invasivo [25].

Questa tesi si inserisce nell'ambito di una collaborazione tra l'Università degli Studi di Padova, il Dipartimento di Informatica dell'Università degli Studi di Milano-Bicocca e l'azienda Seletech Engineering Srl di Brughiero (Milano). La revisione dello stato dell'arte condotta in questo elaborato confronta gli studi riguardanti l'applicazione della spettrometria NIR nel settore agroalimentare in un intervallo temporale che va dal 2010 al 2022. Lo scopo della tesi è di fornire una panoramica sull'evoluzione della tecnica, in particolare sulle principali caratteristiche estraibili dallo spettro elettromagnetico in un intervallo che va da 1350 a 2150 nm e un'analisi sulle principali tecniche di pre-elaborazione e di analisi multivariata.

La presente tesi è strutturata nel seguente modo:

Capitolo 2 – Background: in questo capitolo viene introdotta la spettrometria nel vicino infrarosso come tecnica per l'estrazione di informazioni sulla composizione chimica di prodotti nel settore agroalimentare.

Capitolo 3 – Revisione critica dello stato dell'arte: in questo capitolo viene presentato il lavoro di ricerca bibliografica svolto e una presentazione sulle tecniche di pre-elaborazione dei dati e di analisi multivariata.

Capitolo 4 – Discussione: in questo capitolo viene discussa la letteratura con considerazioni sulle caratteristiche estratte dallo spettro elettromagnetico e sulle principali tecniche di pre-elaborazione e di analisi multivariata utilizzate.

Capitolo 5 – Conclusioni e previsioni future: in questo capitolo sono presentate alcune considerazioni sulla tecnica studiata, sono riassunti i punti di forza e i limiti della tecnica e si discutono le previsioni future.

Capitolo 2

Background

2.1 Introduzione alla spettrometria nel vicino infrarosso

La scoperta dell'energia NIR ebbe luogo a partire dal'800 con gli esperimenti di Herschel ma la sua importanza non venne riconosciuta fino a metà del secolo scorso [1].
Le prime sperimentazioni sulle applicazioni della spettrometria NIR iniziarono a partire dal 1920, ma solamente a partire dal 1960, grazie al contributo di Karl Norris ne venne riconosciuta l'importanza e le sue possibili applicazioni nel settore agroalimentare [1].

Karl Norris lavorò presso il dipartimento dell'Agricoltura degli Stati Uniti dove offrì un contributo fondamentale per lo sviluppo della tecnologia, tale che gli venne attribuito il titolo di "padre della spettrometria nel vicino infrarosso" [2].

Negli ultimi anni il settore agroalimentare, si trova ad affrontare numerose problematiche, tra cui: un incremento della richiesta di prodotti causata dall'aumento esponenziale della popolazione mondiale e la richiesta da parte dei consumatori di garanzie sulla qualità e sulla tracciabilità del prodotto.

I metodi di analisi tradizionali non sono più in grado di soddisfare le esigenze dei consumatori e dei produttori; risultano essere troppo lenti e dispendiosi di energie per un controllo costante e periodico; perciò, sono necessarie nuove tecniche di analisi più efficaci, strumentazioni più accurate e una ricerca di soluzioni innovative e alternative.

La ricerca di nuove tecniche per sostituire quelle tradizionali ha portato allo sviluppo della spettrometria NIR. Una tecnica di analisi che determina le concentrazioni degli elementi chimici che compongono un prodotto in modo rapido e non invasivo. Per permettere un controllo del prodotto in fase di produzione, di lavorazione e anche in fase di imballaggio, assicurando il corretto sviluppo in ogni fase del ciclo produttivo alimentare [4].

Al giorno d'oggi la tecnologia NIR viene ampiamente utilizzata nel settore agricolo e negli allevamenti. Nel settore agricolo la spettrometria NIR viene utilizzata per controllare le proprietà chimiche di un frutto o di una verdura, verificare lo stato del suolo, per prevenire l'insorgere di epidemie [3]. Negli allevamenti la spettrometria NIR viene utilizzata per controllare le feci, verificare la corretta concentrazione dell'*unifeed* e ottenere informazioni sulle caratteristiche del latte [4]. La spettrometria NIR viene anche utilizzata per controllare la qualità di un prodotto finito come ad esempio verificare il contenuto di sale nei prodotti caseari [4]. La spettrometria NIR viene utilizzata per verificare la provenienza del prodotto ed evitare frodi alimentari [3].

La spettrometria NIR è una tecnica di analisi utilizzata per estrarre informazioni sulla qualità dei prodotti. La qualità di un prodotto viene classificata attraverso due categorie di parametri: parametri interni e parametri esterni. I parametri esterni (calibro, colore, imbrunimento, forma) sono osservabili ad occhio nudo, non necessitano di complesse strumentazioni per l'osservazione e rappresentano il primo contatto diretto con il consumatore. I parametri interni (grado zuccherino, durezza, aroma, colore della polpa) non sono facilmente ottenibili, richiedono uno studio più approfondito e sono quelli più interessanti perché offrono informazioni sul sapore, sullo stato di salute e sulla provenienza del prodotto [5].

Per apprendere informazioni sui parametri interni ci sono due metodologie di analisi: l'analisi distruttiva e l'analisi non distruttiva [5]. L'analisi distruttiva consiste nel selezionare il prodotto da studiare ed estrarre dei campioni interni da analizzare in laboratorio. Questo metodo presenta numerosi svantaggi, di cui il principale è l'esigenza di distruggere il campione

rendendolo così invendibile e inutilizzabile provocando uno spreco di materie prime che si traduce in uno spreco di soldi per l'azienda. Inoltre, l'analisi distruttiva non fornisce un'informazione sul singolo prodotto ma un'informazione statistica sulla partita, richiede un pretrattamento dei campioni e l'attesa dei risultati dell'analisi dal laboratorio.

Il metodo non distruttivo permette di ricavare le proprietà chimiche del prodotto in maniera non invasiva, ovvero senza distruggerlo e senza alterarne la composizione. La spettrometria NIR è una metodologia di analisi non distruttiva che fornisce informazioni dettagliate sulla composizione chimica del prodotto in pochi secondi [24] ed è ecosostenibile in quanto non richiede l'utilizzo di reagenti chimici per l'elaborazione dei dati [4].

L'analisi di spettrometria NIR può essere sintetizzata in quattro passaggi:

1) Collezione dei campioni da analizzare
2) Acquisizione degli spettri dei campioni
3) Pre-elaborazione dei dati
4) Analisi multivariata

2.2 Spettro elettromagnetico

La spettrometria nel vicino infrarosso (NIRS) è una tecnica di analisi che utilizza la regione del vicino infrarosso dello spettro elettromagnetico (da 750 nm a 2500 nm).

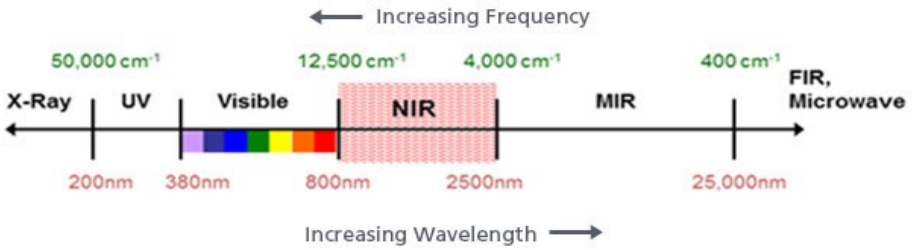

Figura 2.1 Spettro elettromagnetico [6].

Il principio fondamentale su cui si basa la spettrometria NIR è l'osservazione della reazione di alcune molecole particolari all'esposizione a radiazione di diverse lunghezze d'onda nella regione del vicino infrarosso. Le molecole quando vengono irradiate ad una determinata lunghezza d'onda reagiscono alla radiazione ed è possibile identificarle in base alla loro

riflettanza (assorbanza). I legami chimici più interessanti nella regione del vicino infrarosso sono O-H, N-H e C-H. La riflessione registrata produce uno spettro unico per ogni campione che ne rappresenta una "impronta digitale". Lo spettro ottenuto dalla misurazione del campione fornisce, tramite la correlazione con i valori ottenuti dall'analisi in laboratorio, informazioni sulla composizione chimica del campione [11].

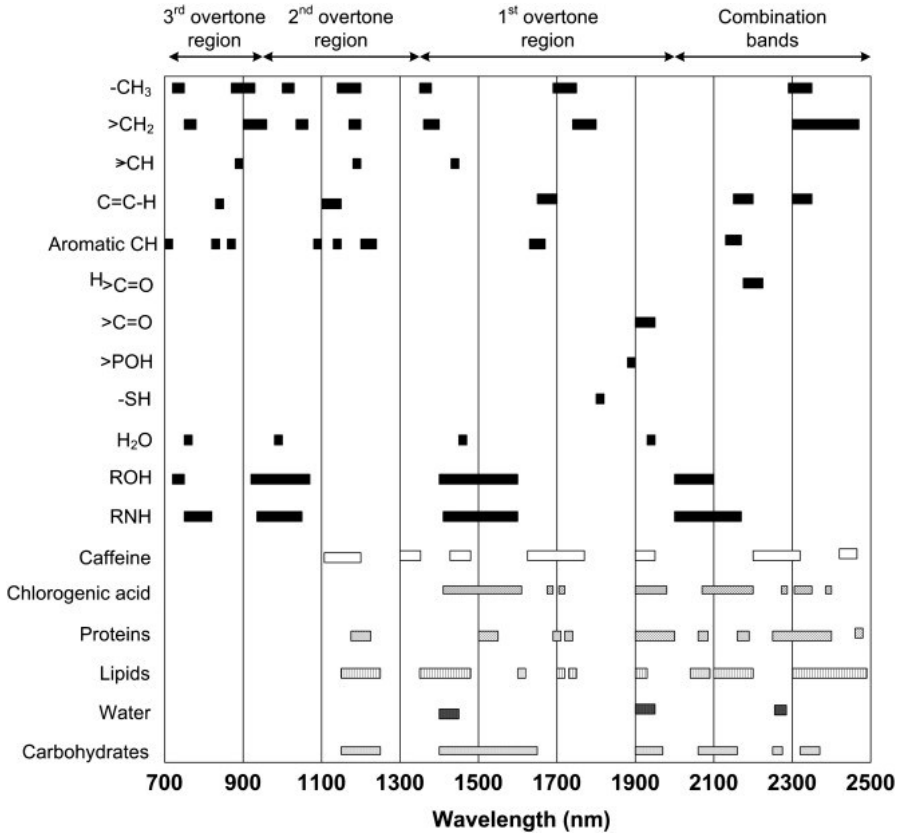

Figura 2.2 Rappresentazione della correlazione tra i principali elementi che compongono il caffè e le lunghezze d'onda [7].

Le misurazioni degli spettri vengono acquisite tramite misure di trasmittanza (*T*) o di riflettanza (*R*) [45]. Nel caso della spettrometria NIR applicata all'estrazione di proprietà di un prodotto si effettuano misure di riflettanza, convertite successivamente in assorbanza (*A*) tramite la formula:

$$A = \log_{10} \frac{1}{R}$$

(1, 1)

2.3 Strumentazione per l'acquisizione dei dati NIRS

Lo strumento utilizzato per l'acquisizione degli spettri è lo spettrometro NIR, uno strumento che produce delle onde elettromagnetiche che vanno a collidere con il campione da studiare, le onde riflesse e/o trasmesse vengono poi analizzate per estrarne le proprietà.

I dispositivi per l'acquisizione degli spettri NIR sono divisibili in due categorie di utilizzo: spettrometri da laboratorio e spettrometri portatili [5].

I dispositivi da laboratorio offrono informazioni più dettagliate ma sono più grandi, costosi e non sono facilmente utilizzabili sul campo.

I dispositivi portatili sono economici e grazie alla loro versatilità possono essere usati in qualsiasi punto del ciclo produttivo. Il principale vantaggio dei dispositivi portatili è la compattezza che li rende tascabili, facilmente trasportabili dagli operatori e utilizzabili in ogni momento. Nonostante le loro ridotte dimensioni mantengono un alto grado di affidabilità, diventando così uno strumento utilissimo nel settore agroalimentare.

Il principio di funzionamento e la composizione di uno spettrometro NIR da laboratorio o portatile è la stessa, cambiano solo alcune componenti hardware.

Lo spettrometro NIR è formato da tre componenti principali: la sorgente, un selettore di lunghezza d'onda e un rilevatore [3].

Figura 2.3 Alcuni esempi di spettrometri NIR portatili: a sinistra, NarutaSpec (Spectral Evolution). Al centro MicroNir (ITPhotonics). A destra Aurora (Grainit).

2.3.1 Sorgente di luce

Le sorgenti di luce utilizzate per gli spettrometri NIR sono tipicamente di due tipi: lampade alogene al tungsteno e diodo a emissione di luce (LED) [3].

Le lampade alogene al tungsteno sono quelle maggiormente utilizzate sia per dispositivi da banco che miniaturizzati, sono affidabili, offrono un'ottima stabilità al raggiungimento dell'equilibrio termico e coprono tutte lunghezze lo spettro del vicino infrarosso, da 300 nm a 2500 nm.

I diodi a emissione di luce (LED) sono molto economici, hanno un basso consumo energetico, sono duraturi e compatti. Raggiungono una lunghezza d'onda di circa 900 nm e per questo non sono propriamente adatti per la spettrometria nel vicino infrarosso.

2.3.2 Selettore di lunghezza d'onda

La selezione d'onda e l'implementazione hardware sono le caratteristiche più critiche di uno spettrometro e ne determinano in gran parte il suo design e i parametri operativi. Le tipologie di selettori di lunghezze d'onda più utilizzati sono l'interferometro Michelson [3], interferometro Fabry-Pérot [3] e la maschera di Hadamard [3].

2.3.3 Rilevatori

I rilevatori (o *photodetector*) servono a registrare le onde riflesse e/o trasmesse attraverso il campione. I principali rilevatori utilizzati nelle strumentazioni NIR sono al silicio, al solfito di piombo (PbS) e all'arseniuro di indio e gallio (InGaAs) [3].

I rilevatori al silicio (Si) sono altamente sensibili fino ad una lunghezza d'onda di 1100 nm, sono veloci, poco rumorosi ed economici. A causa della bassa lunghezza d'onda che possono raggiungere sono più utilizzati nello spettro del visibile (dai 390 ai 700 nm) rispetto al vicino infrarosso [1].

I rilevatori al solfito di piombo (PbS) sono più lenti ma sensibili nella regione che va da 1100 nm a 2500 nm e sono poco diffusi [1].

I rilevatori all'arsenurio di indio e gallio (InGaAs) uniscono i vantaggi del rilevatore al silicio e del rilevatore al solfito di piombo, sono molto sensibili, veloci ma costosi. Lavorano ad una

lunghezza d'onda da 1100 nm a 1600 nm e dispongono di una buona affidabilità anche in un intorno della banda di lavoro mantenendo un buon rapporto segnale-rumore [3].

Oltre alle tipologie citate sono disponibili molti altri modelli di rilevatori ottenuti dalla ricerca di specifiche applicazioni come, ad esempio, i rilevatori InGaAs extended, derivanti dai rilevatori InGaAs, che permettono di arrivare a lunghezze d'onda fino a 2200 nm [3].

La scelta di un solo rilevatore al silicio (Si) è sconveniente per l'analisi NIRS perché le principali informazioni sulle proprietà chimiche si ottengono a lunghezze d'onda maggiori rispetto a quelle rilevabili da questi rilevatori. Per questo solitamente viene adottata una soluzione ibrida. La combinazione dei rilevatori al silicio (Si) e all'arseniuro di gallio (InGaAs) fornisce una soluzione particolarmente favorevole in termini di range, ma ne comporta un aumento dei prezzi di produzione e una diminuzione in termini di compattezza.

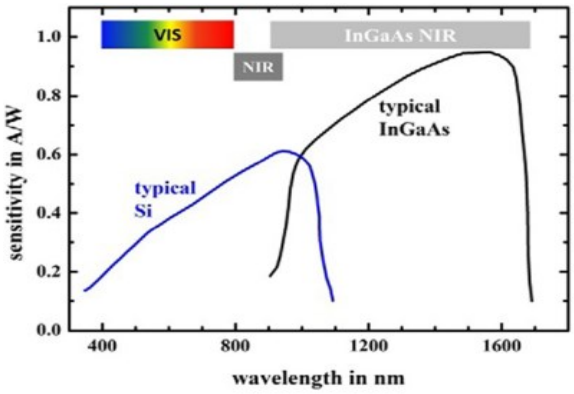

Figura 2.4 Confronto della sensibilità di un rilevatore al silicio (Si) e un rilevatore tipico all'arseniuro di gallio (InGaAs) al variare delle lunghezze d'onda [8].

2.4 Altre tecnologie per l'acquisizione di dati spettrali

La ricerca di soluzioni più efficienti e affabili ha portato allo sviluppo di tecniche e strumentazioni alternative oltre a quelle tradizionali. La spettrometria Vis-NIR può essere applicata a specifici punti del frutto, osservandone piccole aree e producendo uno spettro medio per una specifica superficie, questo tipo di misurazione è detta *points measurament*; oppure, può essere applicata a grandi sezioni del frutto/campione attraverso misure multispettrali e iperspettrali producendo un'immagine della sezione misurata per ciascuna banda di lunghezze d'onda [45]. La differenza tra le misure multispettrali e iperspettrali sta nel numero di bande d'onda utilizzate [45].

L'hyperspectral imagining (HSI) è una delle tecniche di analisi non invasiva più promettenti ed è in continuo sviluppo soprattutto nelle applicazioni nel settore alimentare [12]. Combina la spettroscopia e le tecniche di immagine in un unico sistema per ottenere simultaneamente dati spaziali e spettrali dell'oggetto studiato [12], rivela simultaneamente attraverso speciali rilevatori lo spettro visibile (VIR) e lo spettro nel vicino infrarosso (NIR) associato a ciascun punto dell'immagine.

Figura 2.5: Gaiasky-mini Hyperspectral Imagining system [46].

Capitolo 3

Revisione critica dello stato dell'arte

3.1 Ricerca bibliografica

Per rispondere alle domande di ricerca poste nel primo capitolo si è proceduti a fare una revisione sistematica dello stato dell'arte nell'ambito dell'utilizzo di tecniche di acquisizione e analisi NIRS nel settore agroalimentare.

La tabella 2.1 descrive il lavoro di ricerca bibliografico svolto nella tesi. La tabella consta di venti pubblicazioni ricercate in un intervallo temporale che va dal 2010 al 2022 che utilizzano la spettrometria NIR applicata al settore agroalimentare per estrarre informazioni sulle caratteristiche dei prodotti in modo non invasivo. Nella tabella sono presenti informazioni sul range spettrale, risoluzione spettrale degli strumenti utilizzati nelle pubblicazioni, informazioni sui software per l'acquisizione dei dati (SAD) e sui software per l'analisi statistica (SAS) dei campioni. È indicata la tipologia dell'elemento analizzato (p.e. arance, grano), il numero di campioni studiati e le tecniche di pre-elaborazione dei dati e di analisi multivariata utilizzate. Nel capitolo vengono esposte le principali tecniche di pre-elaborazione dei dati e di analisi multivariata utilizzate nei lavori di ricerca analizzati, con una breve descrizione delle misure di valutazione delle performance dei modelli di regressione e classificazione adottati in letteratura.

Autore e anno di pubbl.	Strumento utilizzato per l'acquisizione degli spettri	Software di acquisizione e/o analisi	Range spettrale [nm] (risoluzione spettrale [nm])	Campione analizzato (feature estratte)	Numerosità dataset [numero e tipologia di campione]	Tecnica di pre-processing più performante (range spettrale in nm)	Metodi di regressione e classificazione	Migliori performance per feature
Martins et al. 2022 [9]	JAZ (Ocean Optics)	SAS: Matlab R2021a (The Mathworks)	680-1100 (-)	Arance (quantità di TSS)	616 arance	-	PLS SpectraNet-23 (algoritmo di deep-learning)	TSS: (R^2=0.31, RMSEP=1.2) TSS: (R^2=0.4, RMSEP=1.16)
Qiao et al. 2022 [10]	Spectrum One NTS FT-NIR (PerkinElmer)	-	680-4800 (3)	Suolo (quantità di Co, Nt)	135 campioni di suolo	-	SVD-CNN SVD-CNN	Co: (R^2=0.8891, RMSE=0.2111) Nt: (R^2=0.9048, RMSE=0.0130)
Heil et al. 2021 [11]	FT-NIRS Vector 22 N (Bruker)	SAS: UnScrambler 10.5 (CAMO Software)	833-2703 (1.2)	Suolo (quantità di Co, Nt, C/N, Cal, ecc)	120 campioni di suolo	Smoothing (S-G derivative, 3 search window) + Derivation (S-G 1° derivata, 1° ordine polinomio) (900-2659)	PLS	Co: (R^2=0.86, RMSEP=0.25%, RPD=10.92) Nt: (R^2=0.98, RMSEP=0.02%, RPD=30.67)
Rabanera et al. 2021 [12]	Hyperspectral imagining system (Specim)	SAD: SpectralDAQ VLNIR (Specim) SAS: ENVI 5.2 + IDL (L3HARRIS), Unscrambler 10.5 (CAMO), Matlab R2013a	898-1750 (3.3)	Noccioli di arachidi (quantità di Um)	200 noccioli di arachidi	-	PLS	Modello costruito utilizzando tutto lo spettro, Um: (R_c^2=0.91, RMSEP=1.87) Modello ottimizzato costruito utilizzando solo delle lunghezze d'onda scelte (R_c^2=0.94, RMSEP=1.95)
Firmani et al. 2020 [13]	FT-NIR Nicolet 6700 (Thermo Scientific Inc)	SAD: Omnic Software (Thermo Scientific) SAS: Matlab R2015b (The Mathworks)	1000-2500 (1)	Pasta (quantità di Pr, Amd, Crb)	949 campioni di pasta	S-G 1° derivata + Mean Center Standard Normal Variate + Mean Centre	Classificazione: PLS-DA Classificazione: SICMA	CC=100% Eff=98.74%, Sn=96.57%, Sp=100%
Borba et al. 2020 [14]	NIR Spectrum 100 N (Perkin-Elmer)	SAS: Origin 8.1 (OriginLab), SAS: Pirouette v4.5 (Infometrix)	1000-2500 (-)	Arance (quantità di V-C, TA, Su, Glu, Fru)	230 arance	Mean center + SNV + S-G 1° derivata (1000-1563) e (1786-2500)	PLS	V-C: (R^2=0.72, RMSEP=107.2) TA: (R^2=0.77, RMSE=4.1) Su: (R^2=0.77, RMSE=11) Glu: (R^2=0.76, RMSE=7.8) Fru: (R^2=0.8, RMSE=5.2)

Le 2020 [15]	M5 NIR (Innovative Technology Group)	SAS: Matlab (The Mathworks)	1100-2490 (2)	Grano (quantità di Um)	80 campioni di grano (dataset disponibile a [16])	-	PLS ELM S-ELM SSAE-AT-ELM	Um: (R^2=0.985, RMSEP=0.0537%) Um: (R^2=0.989, RMSEP=0.0519%) Um: (R^2=0.982, RMSEP=0,0512%) Um: (R^2=0.999, RMSEP=0.0096%)
Firmani et al. 2019 [17]	FT-NIR Nicolet 6700 (Thermo Scientific Inc)	SAS: Omnic Software (Thermo Scientific) SAS: Matlab R2017b (The Mathworks)	1000-2500 (1)	Mandorle	227 mandorle	S-G 1st derivata + Mean Centre Standard Normal Variate + Mean Centre	Classificazione: PLS-DA Classificazione: SICMA	CC=95% Sn=85.7%, Sp=100%
Ni et al. 2019 [18]	MPA II multi-purpose FT-NIR Analyser (Bruker)	SAS: Matlab PLSToolbox (EigenVector)	781-2500 (1)	Pianta di pino (quantità di Nt)	219 foglie di pino	S-G (window size 15, polinomio di 2° ordine)	VWCNN	Nt: (R_p^2=0.925, RMSEP=0.075)
Liu et al. 2018 [19]	Gaia Hyperspectral Imagining System	SAD: ENVI 5.1 SAS: Matlab R2014a (The Mathworks)	863-1704 (2.8)	Foglie di patate (quantità di Um)	71 foglie di patate	-	MLR	Um: (R_c^2=0.95, R_v^2=0.82)
Shi et al. 2017 [20]	Unity SpectraStar 2500XL-R NIR (Unity Scientific)	SAS: UnScrambler 10.4 (CAMO Software)	680-2500 (1)	Grano (quantità di Cp, Um)	80 campioni di grano	Standard Normal Variate + S-G 1° derivata	PLS	Cp: (R_p^2= 0.97, RMSEP=0.31, RPD=6.09)
Bona et al. 2016 [21]	NIRSystem 5000-M (Foss NIR System)	SAD: WinISI III 1.50e (Foss NIRSystem, Tecator Infrasoft International) SAS: Matlab R2008b (The Mathworks) SAS: LIBSVM library v3.20	1100-2498 (2)	Caffè	74 chicchi di caffè	Multiplative Scatter Correction (1100-2498)	Classificazione: SVM	Sn=100%, Sp=100%
Ncama et al. 2016 [22]	NIR Model XDS (Foss NIR System)	SAD: GenStat (VSN International) SAS: Vision (Foss NIR-System)	450-2500 (2)	Arance (quantità di TSS, TA, TSS: TA, BrimA)	120 arance	Nessuna tecnica (856-2292) Standard Normal Variate (856-2292)	PLS	TSS: (R^2=0.747, RMSEP=0.527, RPD=1.73) TSS: (R^2=0.69, RMSEP=0.581, RPD=1.51)
Song et al 2016 [23]	-	SAS: OceanView 2.0 (Ocean Insight)	901-1721 (-)	Mele	60 mele	Baseline Correction + Standard Normal Variate	Classificazione: PLS-DA	CC=98%

Tabella 3.1

Grassi et al. 2014 [24]	MicroNIR 1700 (Diessechem)	SAS: Matlab PLSToolbox (EingenVector)	900-1700 (-)	Caviale (quantità di Um, Pr, Gr)	57 campioni di caviale	Standard Normal Variate + Mean Centre (900-1700)	PLS	Pr: (R^2=0.845, RMSECV=0.515)
Haughey et al. 2012 [25]	FT-NIR Antaris II (Thermo Fisher Scientific)	SAS: TQ Analyst (Thermo Fisher Scientific) SAS: Sicma P+12 (Umetrics)	833-2630 (1.2)	Soia Tostata (quantità di Mel)	-	Standard Normal Variate + S-G 2° derivata (1111-2500)	PLS	Mel: (R^2=0.996, RMSEP=0,158%)
Towett et al. 2012 [26]	Foss 6500 NIR Composite Monochromator (FOSS NIR System)	SAS: WinISI II 1.5 (Infrasoft Int)	-	Fagioli (quantità di Cp)	167 fagioli	- (1100-2498)	M-PLS	Cp: (R_c^2=0.93, R_{cv}^2=0.74)
Sankaran et al. 2011 [27]	SVC HR-1024 (Spectra Vista Cooperation)	SAS: Matlab 7.6 (The Mathworks)	350-2500 (-)	Foglie di agrumi (stato della foglia)	-	-	Classificazione: QDA-SICMA	CC media: 84-87%
Hacisalihoglu et al. 2010 [28]	NIR 256-1.7T1 (Control Development)	SAD: Microsoft Visual Basic 6.0 (Microsoft) SAS: JHP 8 (SAS Institute)	907-1689 (-)	Semi intatti di Fagioli (quantità di Pr, Amd e peso del seme)	273 semi di fagioli	-	PLS	Fw: (R^2=0.85) Pr: (R^2=0.84) Am: (R^2=0.51)
Cayuela et al. 2010 [29]	NIR Labspec (Analitical Spectral Devices) NIR Luminar 5030 (Brimrose)	SAS: Acquire (Brimrose) SAS: UnScrambler 9.7 (CAMO Software)	350-2500 (2) 1100-2300 (2)	Arance (quantità di TSS, Acd, MI, FCI, FW, ecc)	396 arance	Mean Normalization (1100-2300) Mean Normalization + Standard Normal Variate (500-2300)	PLS	TSS: (R^2= 0.92, RMSEP=0.74, RPD=2.13) TSS: (R^2=0.91, RMSEP=0.68, RPD=2.33)

3.2 Metodi di pre-elaborazione dei dati

La pre-elaborazione dei dati, o *pre-processing*, è un passaggio fondamentale nell'analisi chemiometrica [30] e comprende un insieme di tecniche per l'elaborazione dei dati che ha come obiettivo la rimozione di effetti esterni (rumore) indesiderati da un segnale raccolto, per garantire la costruzione di un modello predittivo il più stabile e robusto possibile [31].

Il rumore esterno interferendo con il segnale crea delle alterazioni dei valori che possono non essere riconducibili ad un singolo fattore ma essere generate da errori di misurazione dell'utente, errori strumentali, effetti di scattering e/o altri tipi di rumore [30].

Il passaggio di pre-elaborazione consiste nell'implementazione di uno o più algoritmi per la correzione del segnale e/o l'eliminazione di porzioni dello spettro elettromagnetico caratterizzate da rumore. La selezione del metodo più efficace di pre-processing si rivela quindi essere un passaggio fondamentale e delicato prima del procedere all'analisi di regressione o di classificazione; infatti, applicare uno o più metodi rispetto ad altri, più o meno invasivi, può portare alla rimozione di valori rilevanti per la corretta interpretazione dello spettro e/o lasciare ancora troppi effetti esterni. Per questo a seconda dei dati raccolti e del tipo di analisi effettuata è indispensabile analizzare singolarmente ogni caso. In base al metodo di pre-processing applicato si possono ricavare diversi risultati in termini di performance. Per assicurarsi di aver applicato la tecnica più opportuna è necessaria la sperimentazione e il confronto di più tecniche durante la fase di studio.

Grazie alle molteplici ricerche effettuate nel corso degli anni sono disponibili un gran numero di tecniche di pre-processing. Per comodità si classificano le tecniche di pre-processing in base al loro effetto sullo spettro, ma è possibile trovare differenti classificazioni in quanto non esiste uno standard unico di riferimento ma può variare in base al libro o pubblicazione. *Rinnan et al.* [30] le divide in due categorie: *scatter-correction* e *spectral derivatives*, mentre *Zeng et al.* [31] le divide in quattro categorie: *baseline correction*, *scatter-correction*, *smoothing treatment* e *characteristic scaling*. In questa tesi si farà riferimento alla suddivisione di *Zeng et al. [31]* per una breve descrizione delle loro funzioni e delle più utilizzate tecniche di pre-elaborazione.

3.2.1 Correzione della baseline

Gli algoritmi di pre-processing per la correzione della *baseline* hanno come scopo la rimozione del rumore introdotto dallo strumento che potrebbe causare picchi indesiderati ed errori di calcolo e/o di classificazione [31]. Tra i metodi utilizzati per la *baseline correction* troviamo il *Multiplicative Scatter Correction* (MSC) [22], lo *Standard Normal Variate* (SNV) [22] e il *Savitzky-Golay* (S-G) [31].

3.2.2 Correzione degli effetti di scattering

Gli algoritmi di pre-processing per la *scatter-correction* vengono utilizzati con lo scopo di correggere la dispersione generata da distribuzioni non omogenee del campione e differenti dimensioni delle particelle che causano un aumento della lettura dell'assorbanza.
Tra i più famosi algoritmi di *scatter-correction* ci sono il MSC, l'*Extendend Multiplicative Scatter Correction* (EMSC) [11] e lo *Standard Normal Variate* (SNV) [31].

3.2.3 Correzione del rumore bianco

Gli algoritmi di pre-processing *smoothing treatment* vengono utilizzati per ridurre il rumore bianco. I principali metodi di *smoothing treatment* sono: il *Moving Average*, *Median Filters* e *Savitzky-Golay* [11].

3.2.4 Standardizzazione

Per potere comparare spettri di intensità differenti vengono spesso utilizzati algoritmi di *characteristic scaling* che permettono di scalare lo spettro per compararne le specifiche [11]. Questi metodi includono *mean centering*, *min-max scaling* e la *standardization* [31].

Qui è riportata un breve descrizione di tre importanti algoritmi incontrati nella ricerca bibliografica: *Savitzky-Golay*, il *Multiplicative Scatter Correction* e lo *Standard Normal Variate.*

Savitzky-Golay

L'algoritmo di *Savitzky-Golay* (S-G) è uno dei principali metodi di elaborazione dei dati, è un metodo derivativo, ovvero fa uso delle derivate e il suo obiettivo principale è quello di smussare il segnale ed eliminare effetti additivi e moltiplicativi [30]. Il principio su cui si basa *Savitzky-Golay* è trovare un polinomio di un determinato grado che possa fittare all'interno di un numero di punti scelti (*window size*) del segnale di analisi. Il grado del polinomio e il numero di punti scelti rappresentano parametri di processo; la cui variazione in fase di progettazione dell'algoritmo può fornire risultati con performance molto differenti [30, 31].

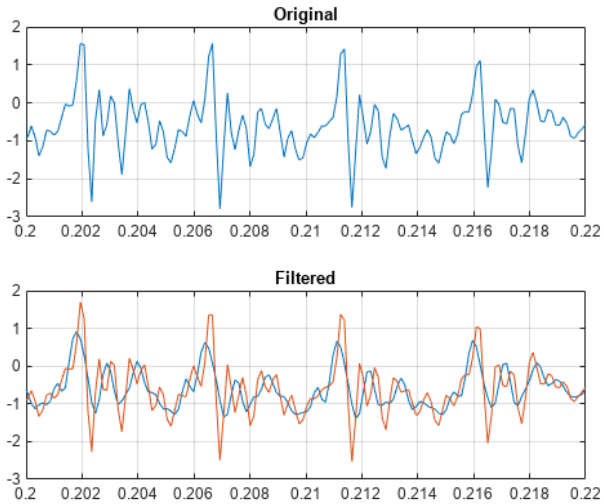

Figura 3.1 Segnale filtrato tramite il filtro Savitzky-Golay con il comando Matlab "sgolayfilt". Figura tratta da [32].

Multiplicative Scatter Correction

Il *Multiplicative Scatter Correction* (MSC) è utilizzato per la correzione degli effetti dispersivi (*scattering*), moltiplicativi e additivi causati dalle differenti dimensioni dei campioni analizzati, dalla morfologia e dall'orientazione delle particelle [21].
Il principio del MSC consiste nel correggere uno spettro misurato rispetto ad uno spettro di riferimento, che può essere la media degli spettri misurati o anche un altro spettro generico. La procedura di pre-processing MSC comprende due passaggi [30]:

1) Stima del coefficiente di correzione.

$$\vec{x}_{org} = b_0 + b_{ref,1} \times \vec{x}_{ref} + \vec{e} \qquad (2, 1)$$

2) Correzione dello spettro.

$$\vec{x}_{corr} = \frac{\vec{x}_{org} - b_0}{b_{ref,1}} = \vec{x}_{ref} + \frac{\vec{e}}{b_{ref,1}}$$

(2, 2)

Dove $\vec{x}_{org}$ è un valore originale dello spettro misurato dallo strumento NIR, $\vec{x}_{ref}$ è lo spettro di riferimento usato per processare l'intero dataset, l'errore $\vec{e}$ è una componente che rappresenta tutti gli effetti dello spettro che non possono essere modellati tramite costanti additive o moltiplicative [39]; $\vec{x}_{corr}$ è lo spettro corretto, b_0 e $b_{ref,1}$ sono costanti che differiscono in base alla tipologia di campione analizzato [30].

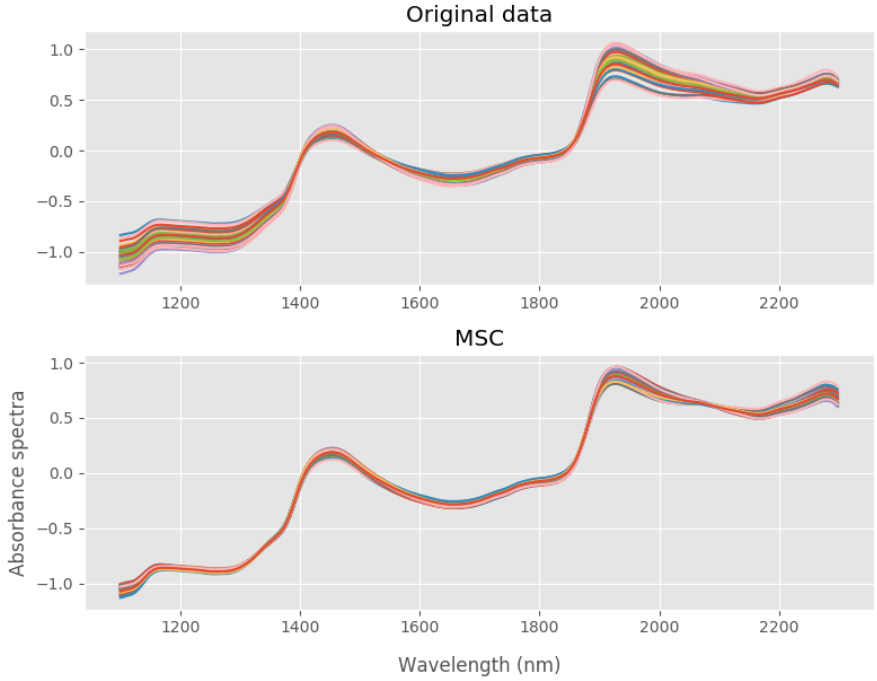

Figura 3.2 Applicazione dell'algoritmo di scattering correction MSC. Figura tratta da [34].

Standard Normal Variate

Lo *Standard Normal Variate* (SNV) viene utilizzato per la correzione degli effetti dispersivi (*scattering*). Il concetto di correzione del segnale SNV è lo stesso del MSC riportato in precedenza, ad eccezione del segnale di riferimento che nel caso del SNV non è richiesto. Per la gran parte delle applicazioni i risultati del SNV e del MSC sono analoghi [30, 33].

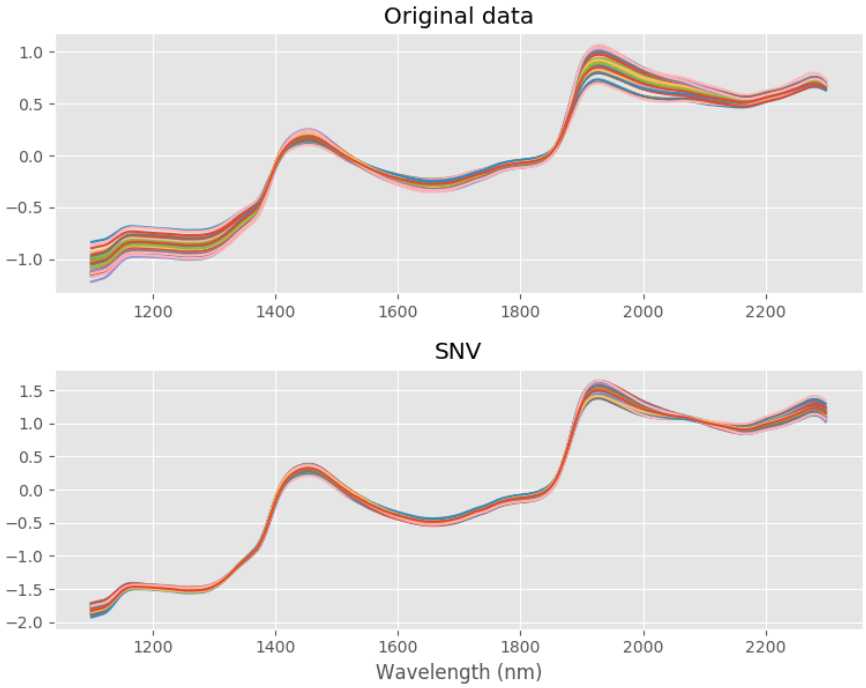

Figura 3.3 Applicazione dell'algoritmo di scattering correction SNV. Figura tratta da [34].

3.3 Metodi di analisi multivariata

L'analisi multivariata, o *multivariate analysis* (MVA), si riferisce allo studio di sistemi con due o più variabili e viene applicata dopo il passaggio di pre-processing, ovvero dopo aver rimosso gli effetti rumorosi dal segnale. La scelta della tecnica di analisi multivariata è anch'esso un passaggio delicato, ma necessario per costruire un modello di previsione affidabile che permetta una corretta predizione delle proprietà per i campioni al di fuori del dataset in cui si sta eseguendo la calibrazione.

Nel corso degli anni sono stati sviluppati numerosi modelli di analisi multivariata con lo scopo di estendere il campo di utilizzo e di migliorare l'affidabilità dei modelli costruiti.

I modelli di analisi multivariata possono essere divisi in metodi di semplificazione dei dati, di regressione e di classificazione. Il principale modello per la semplificazione dei dati è la *Principal Component Analysis* (PCA) [31]. Tra i modelli di regressione più utilizzati ci sono la *Partial Least Square* (PLS) [33], la *Multiple Linear Regression* (MLR) e la *Principal Component Regression* (PCR). I modelli di classificazione hanno un ulteriore sottodivisione in algoritmi supervisionati (*supervised*) e non supervisionati (*unsupervised*). Tra i modelli supervisionati più utilizzati ci sono la *Partial Least Square Discriminant Analysis* (PLS-DA) [24], il *Support Vector Machine* (SVM) [22], il *Random Forest (RF) [11],* il K-Nearest Neighbor (K-NN) *[8]* e il *Soft Independent Modeling of Class Analogies* (SICMA) [13]. Mentre tra i modelli non supervisionati più utilizzati troviamo il *Convolutional Neural Network* (CNN) [10] e l'*Extreme Learning Machine* (ELM) [15].

3.3.1 Modelli di semplificazione dei dati

I modelli di semplificazione dati vengono utilizzati per fare un'analisi esplorativa dei dati a disposizione, per ridurre la dimensione delle variabili e per rimuovere misurazioni anomale ottenute *(outliers)*, in modo da evitare di costruire un modello alterato da poche misurazioni anomale.

Principal Component Analysis (PCA)

La PCA, o analisi delle componenti principali è una tecnica per la semplificazione dei dati e rappresenta uno strumento importante come primo approccio al trattamento di dati complessi. La PCA è utilizzata come analisi esplorativa e mi permette di visualizzare i dati correlati e ridurre il numero di variabili che descrivono un problema con la minore perdita di informazioni possibile [33].

3.3.2 Modelli di regressione

I modelli di regressione vengono utilizzati per la costruzione di modelli quantitativi al fine di predire la composizione chimica di un campione a partire dai dati spettrali raccolti [25].
La creazione di un modello di regressione richiede tre passi fondamentali: la fase di addestramento o di calibrazione (*calibration o training*), la fase di validazione incrociata (*cross-validation)* e la fase di predizione (*prediction o test*). Sono disponibili numerosi modelli di regressione come la regressione PL, la PCR e la MLR.

Partial Least Square (PLS)

La PLS, o regressione ai minimi quadrati parziali è il modello di regressioni più preponderante per le analisi multivariate in questo contesto applicativo. Secondo la revisione dello stato dell'arte condotto da *Dotto et al.* [33] nel 2018, tra il 2006 e il 2016 la regressione PLS compariva come metodo di analisi multivariata per la predizione delle proprietà del suolo con una frequenza del 65%. La PLS mette in relazione le informazioni presenti su due tabelle di dati dello stesso insieme di osservazione [35]. Nel nostro caso mette in relazione gli spettri misurati con la strumentazione NIR con i valori ottenuti tramite le misurazioni da laboratorio tradizionali con l'obiettivo di trovare un legame tra i due.

3.3.3 Valutazione delle performance di un modello di regressione

La necessità di valutare le performance di un modello di regressione per confrontarne le prestazioni richiede l'introduzione di alcuni parametri. Per la valutazione della bontà del

modello si utilizzano principalmente tre coefficienti: il *coefficiente di determinazione* (R^2), la *radice dell'errore quadratico medio* (RMSE) e la *deviazione dei residui predetti* (RPD) [11].

R^2, o *coefficient of determination*, è un indice che fornisce informazioni sulla bontà del fitting del modello [11] e rappresenta la misura di gran lunga più utilizzata per la valutazione dei modelli [33]. Il coefficiente di determinazione può essere ricavato con la formula:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

(2, 3)

dove $\bar{y}$ è la media dei valori osservati, $\hat{y}$ è il valore predetto, y è il valore osservato e n è il numero totale di dati analizzati. Esso può essere riferito alla fase di *training* (R_c^2), alla fase di cross-validazione (R_{cv}^2) o alla fase di predizione (R_p^2).
Il risultato dell'eq. (2, 2) fornisce un numero intero adimensionale compreso tra 0 e 1; maggiore è il valore R^2 trovato e maggiore è l'affidabilità del modello.

IL RMSE, o *Root Mean Square Error*, è un indice dell'accuratezza dei calcoli [11], misura la differenza tra i valori calcolati e i valori osservati dal modello misurato e può essere calcolata con la formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{n}}$$

(2, 4)

dove $\bar{y}$ è la media dei valori osservati, $\hat{y}$ è il valore predetto e n è il numero totale di dati analizzati. L'RMSE è un numero sempre maggiore di 0 e dimensionale: questo rappresenta un particolare limite di questo parametro che lo rende difficilmente confrontabile con modelli differenti. Per la costruzione di un modello accurato si cerca il minor valore di RMSE possibile. Analogamente al R^2, possiamo trovare l'RMSE riferito alla fase di calibrazione (RMSEC), alla fase di cross-validazione (RMSECV) o alla fase di predizione (RMSEP).

La RPD, o *ratio of the performance deviation*, è un indice dell'accuratezza dei modelli [11] ed è calcolabile con la formula:

$$RPD = \frac{SD}{RMSE}$$

(2, 5)

dove SD indica la deviazione standard, del campione di osservazione a disposizione. L'RPD è sempre un numero maggiore di 0 e adimensionale; quindi, per la costruzione di modelli si cerca di trovare il valore di RPD maggiore. Secondo *Ncama et al.* [22], un modello accettabile ha un valore RPD compreso tra 1.4 e 2, mentre un modello con RPD maggiore di 2 è da considerarsi buono.
Secondo *Heil et al.* [11], nella valutazione delle tecniche di pre-processing per il trattamento del suolo, un valore di RPD pari a 4.1 è considerato eccellente mentre 1.9 è la soglia minima accettabile per garantire una buona accuratezza del modello in quel contesto applicativo.

Oltre agli indici di valutazione introdotti sopra sono disponibili altri indici importanti ma meno utilizzati nella letteratura, come la media della differenza tra i valori misurati e calcolati (Bias) [23] e l'errore standard di predizione (SEP) [11].

3.3.4 Modelli di classificazione

I modelli di classificazione sono uno strumento importante per l'analisi qualitativa dei campioni e hanno l'obbiettivo di classificare correttamente un campione all'interno di un dataset come appartenente ad una specifica classe o condizione.

Per confrontare la bontà di un modello di classificazione è necessario, come nei modelli di regressione, introdurre alcune misure di performance che servono per il confronto con altri modelli. I parametri maggiormente utilizzati sono tre: l'accuratezza (*accuracy)*, la sensibilità (*sensitivity*) e la specificità (*specificity*), solitamente espressi in valori percentuali [36].

Nella revisione dello stato dell'arte effettuata in questo lavoro di tesi, sono state ottenute tali misure a partire dai seguenti valori: il numero di veri positivi, o *true positive* (TP), il numero di falsi positivi, o *false positive* (FP), il numero di veri negativi, o *true negative* (TN) e il numero di falsi negativi, o *false negative* (FN) [36].

Il coefficiente di accuratezza chiamato anche coefficiente di corretta classificazione, o *accuracy/correct classification*, è utilizzato come indice di quanto il modello abbia predetto correttamente i valori studiati. L'accuratezza si calcola come il numero di predizioni corrette diviso per il numero totale di predizioni, e viene calcolata con la formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(2, 6)

Il coefficiente di sensibilità, o *sensitivity*, è utilizzato come indice della precisione con cui i campioni della classe "positivi" sono stati correttamente etichettati come "positivi" e viene calcolato con la formula:

$$Sensitivity = \frac{TP}{TP + FN}$$

(2, 7)

Il coefficiente di specificità, o *specificity*, fornisce un indice di corretta classificazione dei campioni "negativi" e viene calcolato con la formula:

$$Specificity = \frac{TN}{FP + TN}$$

(2, 8)

Capitolo 4

Discussione

4.1 Considerazioni sulle tecniche di pre-processing

Per valutare le performance dei modelli di regressione e di classificazione è necessario stabilire una soglia per capire se il risultato ottenuto è sufficiente o insufficiente. Dopo aver revisionato numerose pubblicazioni è possibile constatare che per mantenere un alto grado di affidabilità, si può fissare un valore di R^2 a 0.9 per considerare un modello sufficiente, per valori superiori a 0.95 valutiamo un modello ottimo e per valori inferiori allo 0.9 lo giudichiamo insufficiente. Per i modelli di classificazione, valutiamo un modello ottimo per valori sopra al 95%, sufficiente per valori superiori al 85% e insufficiente per valori inferiori al 85%.

La revisione della letteratura per quanto riguarda le tecniche di pre-processing ha portato alla conclusione che non esiste una tecnica o più tecniche che combinate possano essere considerate ottimali per qualsiasi dataset. La scelta della tecnica o delle tecniche più opportune sono decisioni da compiere di volta in volta.
Tuttavia, nella ricerca bibliografica svolta, sono state riscontrate alcune tecniche che hanno ottenuto delle ottime performance per gran parte dei modelli costruiti.

La tecnica derivativa *Savitzy-Golay* ha fornito buoni risultati nella maggior parte delle pubblicazioni in cui è stata applicata. Ad esempio, nella pubblicazione di *Heil et al.* [11] dove sono state confrontate cinquantacinque tecniche di pre-processing differenti per valutare proprietà del suolo, tra tutte *Savitzky-Golay* ha ottenuto i coefficienti di determinazione più alti e i valori di RMSE più bassi. Nella revisione di *Rinnan et al.* [29] per l'analisi di campioni di marzapane viene riscontrato che a basse lunghezze d'onda è sufficiente applicare la *normalization,* mentre per tutti gli altri casi MSC e SNV si possono utilizzare indifferentemente perché forniscono circa le stesse performance.

4.2 Riflessione sull'analisi dello spettro elettromagnetico e sui modelli di analisi multivariata

Dopo il pre-processing per estrarre le informazioni dai dati è necessario scegliere un modello di analisi multivariata. Il modello di regressione più utilizzato nelle pubblicazioni revisionate è il *Partial Least Square* e il parametro più utilizzato per descriverne le performance è il coefficiente di determinazione.

Per quanto riguarda lo studio dello spettro, ovvero la scelta della lunghezza d'onda e della banda da esaminare, è da analizzare ogni caso singolarmente. Nella tabella 3.1 sono riassunte alcune lunghezze d'onda importanti estratte dalla revisione della letteratura. Per le pubblicazioni visionate lo spettro elettromagnetico nella regione da 1350 a 2150 nm permette di estrarre informazioni dai campioni sul contenuto d'acqua, di carboidrati, di carbonile e di proteine, oltre a fornire un'impronta digitale unica del campione analizzato. Lo spettro può essere utilizzato come timbro per distinguere una gamma differente di campioni e/o per trovare campioni adulterati. Ad esempio, nella pubblicazione di *Hacisalihoglu et al.* [28], si notano come le varietà di fagioli neri hanno un picco di assorbanza nella regione tra 910 e 1100 nm. Questo picco non è stato osservato in altre colorazioni di fagioli e permette, grazie al confronto degli spettri di distinguere varie tipologie di fagioli da quelli neri. Nella pubblicazione di *Haughey et al.* [25] vengono contaminati dei campioni di soia con della melanina per creare dei modelli contraffatti e grazie al confronto dello spettro normale con quello alterato (Figura 3.1) è possibile estrarre informazioni sulla presenza di melanina e identificare i modelli alterati.

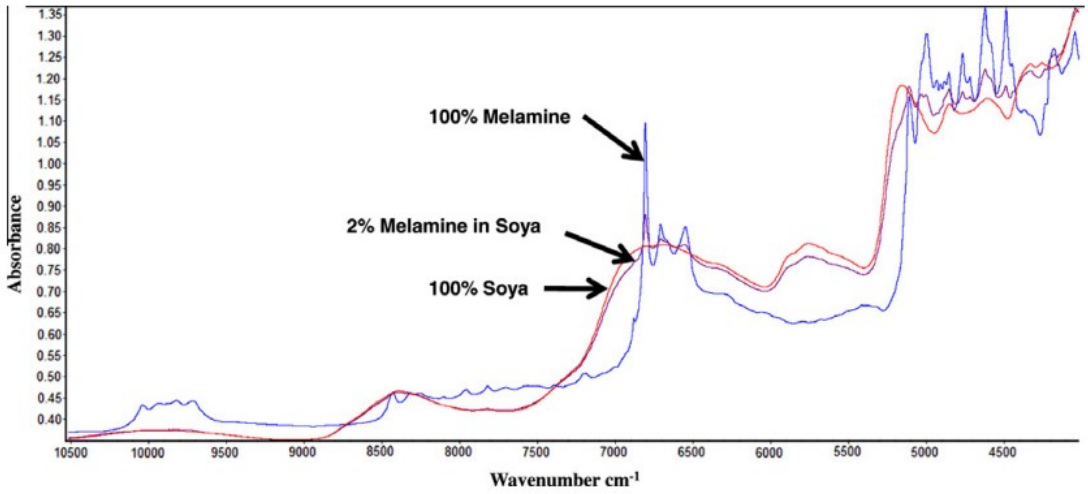

Figura 4.1 Tratta da [22].

Anche nella pubblicazione di *Firmani et al.* [13] si ottengono buoni risultati. Sono stati raccolti gli spettri di campioni di pasta Gragnano e non Gragnano e tramite la costruzione di un modello di classificazione PLS-DA, si riesce a identificare con un'accuratezza del 100% i campioni di pasta Gragnano e con un'accuratezza del 98.10% i campioni di pasta non Gragnano.

Elemento analizzato	Pubblicazione	Lunghezze d'onda [nm] e feature estratte
Arance	[6, 26]	680 – contenuto solido solubile 1400-1500, 1880-2100 - umidità 1850 - carboidrati
Suolo	[8]	1420-1730, 1840-2060, 2160-2600 – carbonio, contenuto organico 1490-1730, 1850-1970, 2000-2100, 2149 - carbonile
Pasta	[10]	2500 – proteine, amido, carbonile 1429 - carboidrati
Grano	[17]	1400-1470, 1900-1950 – umidità 2148-2200 - proteina
Soia (adulterata con melanina)	[22]	1480 – melanina

Tabella 4.1

Il valore con cui decidere se un modello di classificazione fornisce dei risultati insufficienti, sufficienti o ottimi resta comunque da stabilire. Un valore di classificazione prossimo al 100% può non essere comunque sufficiente. Nel caso di *Wang et al.* [40], dove vengono analizzati dei campioni di porcini, il margine d'errore non esiste, la classificazione sbagliata rischierebbe di etichettare un fungo infetto come sano mettendo a rischio la vita del consumatore. È necessario quindi anche in questa dinamica, valutare di volta in volta quali siano i valori di corretta classificazione accettabili. Nel resto delle pubblicazioni analizzate [14, 18, 20, 21, 24] dove non sorgono questi problemi, si trovano risultati promettenti per la classificazione corretta dei prodotti con coefficienti di accuratezza maggiore del 84%.

Capitolo 5

Conclusioni e previsioni future

L'attività di ricerca svolta in questa tesi mostra come la spettrometria nel vicino infrarosso (NIRS) sia uno strumento utile per l'analisi dei prodotti nel settore agroalimentare. Pur essendo una tecnica di analisi relativamente nuova, è ormai ampiamente consolidata e adottata da moltissime industrie nel settore.

Si è visto, come la spettrometria NIR consenta un'analisi accurata e non invasiva del prodotto con i risultati disponibili dopo pochi secondi, senza necessità di particolari pretrattamenti sui campioni da studiare e con la possibilità di effettuare le misurazioni direttamente sul campo grazie alla strumentazione portatile. Inoltre, permette l'abbattimento dei costi per l'analisi e un risparmio di reagenti per lo studio in laboratorio e per il trattamento dei campioni.

Tuttavia, la strada da percorrere per considerare la spettrometria NIR come una tecnica di analisi completa è ancora lunga. Al giorno d'oggi ci sono ancora numerose problematiche a cui far fronte, tra cui la principale è la costruzione di modelli di predizione affidabili e robusti. Nella revisione degli articoli fatta in questa tesi si è visto come l'applicazione di una o più tecniche di pre-elaborazione o di analisi multivariata rispetto ad altre impatti notevolmente le performance del modello, per cui durante la progettazione è necessario procedere analizzando caso per caso e confrontandone i risultati. Inoltre, la realizzazione di modelli necessita di grandi

banche di dati da correlare tra le misure spettrali acquisite e tra i campioni analizzati in laboratorio, operazione che richiede tempo e personale qualificato.

Allo stato attuale, ci si interroga su quali siano le possibili soluzioni per far fronte alle sue attuali limitazioni. Molte pubblicazioni si affidano all'applicazione di alcune tecniche innovative che hanno già offerto ottimi risultati in altri settori come, per esempio, gli algoritmi di *machine learning,* già utilizzati nel settore sanitario per la realizzazione di tecnologia indossabile per il rilevamento di eventi dell'andatura [43] e la classificazione dei vari compiti motori [44]. Queste tecniche, di cui si è fatta visione anche nelle pubblicazioni all'interno della tesi, risultano essere molto performanti ma introducono problemi computazionali dovuti alla grande mole di dati da analizzare.

Questa tesi si inserisce nell'ambito di una collaborazione tra l'Università degli Studi di Padova, il Dipartimento di Informatica dell'Università degli Studi di Milano-Bicocca e l'azienda Seletech Engineering Srl di Brughiero (Milano).

Le prospettive per il futuro della spettrometria nel vicino infrarosso rivelano come sia necessario ancora un grande lavoro di ricerca per portare questa tecnologia all'apice delle sue possibilità. Il numero di pubblicazioni crescenti negli ultimi decenni e le moltissime aziende che si stanno mobilitando per produrre strumentazioni NIRS, sono indice di come questa nuova tecnologia abbia attirato l'interesse dei ricercatori e delle industrie del settore.

Bibliografia:

[1] Nicola Berardo, S. Locatelli, *Applicazione della spettrometria nel vicino infrarosso (NIR) nel settore agroalimentare*. AIPnD Congresso, 2009.

[2] Williams P. Karl H. Norris, *the Father of Near-Infrared Spectroscopy*. NIR news, vol. 30, 25-27, 2019.

[3] K. B. Beć, J. Grabska, and C. W. Huck, *Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives*. Foods, vol. 11, no. 10, p. 1465, 2022.

[4] C. Evangelista, L. Basiricò, and U. Bernabucci, *An Overview on the Use of Near Infrared Spectroscopy (NIRS) on Farms for the Management of Dairy Cows*. Agriculture, vol. 11, no. 4, p. 296, 2021.

[5] Vincenzo Girgenti, C. Peano, Nicole Giuggioli, *L'utilizzo della spettrofotometria NIR nella valutazione non distruttiva delle caratteristiche qualitative dei frutti.* AIPnD Congresso, 2009.

[6] Figura 2.1 disponibile al link: https://www.fossanalytics.com/en/news-articles/technologies/nir-technology (accesso il 2 novembre 2022).

[7] Douglas Fernandes Barbin, Ana Lucia de Souza Madureira Felicio, Da-Wen Sun, Suzana Lucy Nixdorf, Elisa Yoko Hirooka, *Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview*. Food Research International, vol 61, 23-32, 2014.

[8] Figura 2.3 disponibile al link: https://photinnov.com/en/quest-ce-que-lingaas/ (accesso il 2 novembre 2022).

[9] J.A. Martins, R. Guerra, R. Pires, M.D. Antunes, T. Panagopoulos, A. Brázio, A.M. Afonso, L. Silva, M.R. Lucas, A.M. Cavaco, *SpectraNet–53: A deep residual learning architecture for*

predicting soluble solids content with VIS–NIR spectroscopy. Computers and Electronics in Agriculture, vol. 197, 2022.

[10] Hanli Qiao, Xiubo Shi, Huazhou Chen, Jingyi Lyu, Shaoyong Hong, *Effective prediction of soil organic matter by deep SVD concatenation using FT-NIR spectroscopy.* Soil and Tillage Research, vol. 215, 2022.

[11] K. Heil, U. Schmidhalter, "An Evaluation of Different NIR-Spectral Pre-Treatments to Derive the Soil Parameters C and N of a Humus-Clay-Rich Soil," *Sensors*, vol. 21, no. 4, p. 1423, 2021.

[12] J.D. Rabanera, J.D. Guzman, K.F. Yaptenco, *Rapid and Non-destructive measurement of moisture content of peanut (Arachis hypogaea L.) kernel using a near-infrared hyperspectral imaging technique*. Food Measure 15, 3069–3078, 2021.

[13] Patrizia Firmani, Giuseppe La Piscopia, Remo Buccia, Federico Marinia, Alessandra Biancolillo, *Authentication of P.G.I. Gragnano pasta by near infrared (NIR) spectroscopy and chemometrics.* Microchemical Journal, vol. 152, 2020.

[14] Karla Borba, Poliana Spricigo, Didem Aykas, Milene Mitsuyuki, Luiz Colnago, Marcos Ferreira. Non-invasive quantification of vitamin C, citric acid, and sugar in 'Valência' oranges using infrared spectroscopies. Journal of Food Science and Technology, vol. 58, 2020.

[15] Ba Tuan Le, *Application of deep learning and near infrared spectroscopy in cereal analysis.* Vibrational Spectroscopy, vol. 106, 2020.

[16] Dataset del grano disponibile al link: https://eigenvector.com/resources/data-sets/ (ultimo accesso 1° novembre 2022).

[17] Patrizia Firmani, Remo Bucci, Federico Marini, Alessandra Biancolillo, *Authentication of "Avola almonds" by near infrared (NIR) spectroscopy and chemometrics.* Journal of Food Composition and Analysis, vol. 82, 2019.

[18] Chao Ni, Dongyi Wang, Yang Tao, *Variable weighted convolutional neural network for the nitrogen content quantization of Masson pine seedling leaves with near-infrared spectroscopy.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 209, 32-39, 2019.

[19] Ning Liu, Li Wu, Longsheng Chen, Hong Sun, Qiaoxue Dong, Jingzhu Wu, *Spectral Characteristics Analysis and Water Content Detection of Potato Plants Leaves*. IFAC-PapersOnLine, vol. 51, 541-546, 2018.

[20] Haitao Shi, Peiqiang Yu, *Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat.* Food Control, vol. 82, 57-65, 2017.

[21] Evandro Bona, Izabele Marquetti, Jade Varaschim Link, Gustavo Yasuo Figueiredo Makimori, Vinícius da Costa Arca, Andre Luis Guimar aes Lemes, Juliana Mendes Garcia Ferreira, Maria Brígida dos Santos Scholz, Patrícia Valderrama, Ronei Jesus Poppi, *Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee*. Food Science and Technology, vol. 76, Part B, 330-336, 2017.

[22] Khayelihle Ncama, Umezuruike Linus Opara, Samson Zeray Tesfay, Olaniyi Amos Fawole, Lembe Samukelo Magwaza, *Application of Vis/NIR spectroscopy for predicting sweetness and flavour parameters of 'Valencia' orange (Citrus sinensis) and 'Star Ruby' grapefruit (Citrus x paradisi Macfad).* Journal of Food Engineering, vol. 193, 86-94, 2017.

[23] Weiran Song, Hui Wang, Paul Maguire, Omar Nibouche, *Differentiation of Organic and Non-organic Apples Using Near Infrared Reflectance Spectroscopy – A Pattern Recognition Approach.* 2016 IEEE Sensors, 1-3, 2016.

[24] Maurizio Grassi, Stefania Barzaghi, Marina Buccheri, Mario Pazzaglia, Mauro Vasconi, Tiziana M.P. Cattaneo, *La spettroscopia NIR applicata al controllo qualità del caviale: risultati preliminari. Intervento presentato al sesto convegno Simposio italiano di spettroscopia NIR tenutosi a Modena nel 2014,* 2014.

[25] Simon A. Haughey, Stewart F. Graham, Emmanuelle Cancouët, Christopher T. Elliott, *The application of Near-Infrared Reflectance Spectroscopy (NIRS) to detect melamine adulteration of soya bean meal.* Food Chemistry, vol. 136, 1557-1561, 2013.

[26] E.K., Towett, M. Alex, K.D. Shepherd, S. Polreich, E. Aynekulu, B.L. Maass, Applicability of near-infrared reflectance spectroscopy (NIRS) for determination of crude protein content in cowpea (*Vigna unguiculata*) leaves. Food Sci Nutr, 45-53, 2013.

[27] Sindhuja Sankaran, Reza Ehsani, *Visible-near infrared spectroscopy based citrus greening detection: Evaluation of spectral feature extraction techniques*. Crop Protection, vol. 30, 1508-1513, 2011.

[28] Gokhan Hacisalihoglu, Bismark Larbi, A. Mark Settles, *Near-Infrared Reflectance Spectroscopy Predicts Protein, Starch, and Seed Wight in Intact Seeds of Common Bean (Phaseolus vulgaris L.).* Journal of Agricultural and Food Chemistry, 702-706, 2010.

[29] José A. Cayuela, Carlos Weiland, *Intact orange quality prediction with two portable NIR spectrometers.* Postharvest Biology and Technology, vol. 58, 113-120, 2010.

[30] Asmund Rinnan, Frans van den Berg, Søren Balling Engelsen, *Review of the most common pre-processing techniques for near-infrared spectra.* Trends in Analytical Chemistry, vol. 28, no. 10, 2009.

[31] Jian Zeng, Yuan Guo, Yanqing Han, Zhanming Li, Zhixin Yang, Qinqin Chai, Wu Wang, Yuyu Zhang, Caili Fu, *A Review of the Discriminant Analysis Methods for Food Quality Based on Near-Infrared Spectroscopy and Pattern Recognition. Molecules*, vol. 26, no. 3, p. 749, 2021.

[32] Figura 2.1 disponibile al link: https://it.mathworks.com/help/signal/ref/sgolayfilt.html (accesso il 2 novembre 2022).

[33] Andre Carnieletto Dotto, Ricardo Simao Diniz Dalmolin, Alexandre ten Caten, Sabine Grunwald, *A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra.* Geoderma, vol. 314, 262-274, 2018.

[34] Figura 3.2-3.3 disponibile al link: https://nirpyresearch.com/two-scatter-correction-techniques-nir-spectroscopy-python/ (accesso il 2 novembre 2022).

[35] Herve´ Abdi, Lynne J. Williams, *Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression.* Computational Toxicology, vol. 930, 549-579, 2012.

[36] Wikipedia, L'enciclopedia libera, *Sensitivity and specificity*, disponibile al link: https://en.wikipedia.org/wiki/Sensitivity_and_specificity#Specificity (accesso il 2 novembre 2022).

[37] M. Deepak, S. Keski-Saari, L. Fauch, L. Granlund, E. Oksanen, M. Keinänen, *Leaf Canopy Layers Affect Spectral Reflectance in Silver Birch*. Remote Sensing, vol. 11, no. 24, p. 2884, 2019.

[38] Joel B Johnson, *Discrimination of centre composition in panned chocolate goods using near infrared spectroscopy*. Journal of Near Infrared Spectroscopy, 30, 3, 130-137, 2022.

[39] M.R. Maleki, A.M. Mouazen, H. Ramon, J. De Baerdemaeker, *Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy*. Biosystems Engineering, vol 96, 3, 427-433, 2007.

[40] Li Wang, Honggao Liu, Jieqing Li, Yuanzhong Wang, *Verified the rapid evaluation of the edible safety of wild porcini mushrooms, using deep learning and PLS-DA.* Journal of the Science of Food and Agriculture, vol 102, 4, 1531-1539, 2022.

[41] Angelita Rettore de Araujo Zanella, Eduardo da Silva, Luiz Carlos Pessoa Albini, *Security challenges to smart agriculture: Current state, key issues, and future directions*. Array, vol 8, 2020.

[42] A. Zancanaro, G. Cisotto and L. Badia, *Challenges of the Age of Information Paradigm for Metrology in Cyberphysical Ecosystems*. 2022 IEEE International Workshop on Metrology for Living Environment (MetroLivEn), 127-131, 2022.

[43] Matteo Gadaleta, Giulia Cisotto, Michele Rossi, Rana Zia Ur Rehman, Lynn Rochester, Silvia Del Din, *Deep Learning Techniques for Improving Digital Gait Segmentation*. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021.

[44] Giulia Cisotto, Anna V. Guglielmi, Leonardo Badia and Andrea Zanella, *Classification of grasping tasks based on EEG-EMG coherence*. 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 1-6, 2018.

[45] Ana Cavaco, Dário Passos, Rosa Pires, Maria Antunes, Rui Guerra, *Nondestructive Assessment of Citrus Fruit Quality and Ripening by Visible-Near Infrared Reflectance Spectroscopy. Citrus - Research, Development and Biotechnology. London, United Kingdom: IntechOpen,* 2021.

[46] Figura 2.5 tratta da: http://www.zolix.com.cn/en/Product_desc/1240_1530.html (accesso il 12 novembre 2022).